The Journal of Experimental Education

A periodical report of scientific investigations relating to child development, curriculum, learning, teaching, supervision, measurements, statistics, and experimental techniques.

lume XXIV September, 1955 Number 1

CONTENTS

.00 A YEAR PUBLISHED QUARTERLY $1.50 A COPY

Published by Dembar Publications, Inc.,
Madison 3, Wisconsin.
Entered as second-class matter October 17, 1938 at the post office at Madison, Wisconsin, under the act of March 3, 1879.

EDITORIAL BOARD

CONTRIBUTING EDITORS

DEMOCRAT PRINTING COMPANY
MADISON, WISCONSIN

The Journal of Experimental Education

A periodical report of scientific investigations relating to child development, curriculum, learning, teaching, supervision, measurements, statistics, and experimental techniques.

Volume XXIV | December, 1955 | Number 2

CONTENTS

$5.00 A YEAR | PUBLISHED QUARTERLY | $1.50 A COPY

Published by Dembar Publications, Inc.,
Madison 3, Wisconsin.
Entered as second-class matter October 17, 1938 at the post office at Madison, Wisconsin, under the act of March 3, 1879.

EDITORIAL BOARD

CONTRIBUTING EDITORS

DEMOCRAT PRINTING COMPANY
MADISON, WISCONSIN

Journal of Experimental Education

Volume XXIV September, 1955 Number 1

A STUDY OF PROFESSIONAL DISTANCES BETWEEN THE RATERS OF TEACHERS AND TEACHERS RATED*

EARL MARTIN GROTKE
University of Southern California

SECTION I

The Problem

THIS STUDY attempts to show the relationship between the attitudes of raters and ratees and the ratings given to teachers. It is hypothesized that the lengths of "professional distances" between raters and ratees increase as teacher ratings decrease from good to average and average to poor.

Definition of Professional Distance.—The concept of professional distance is adapted from the concept of social distance used in the field of sociology. In 1925, while studying racial attitudes, Bogardus devised an attitude scale called social distance. "He was interested in measuring degrees to which individual representatives of various racial and national groups were accepted or rejected.... Instead of making a distinction between favorable and unfavorable attitudes, however, he conceived the problem in terms of degrees of 'distance' which his subjects wished to keep between themselves and members of other groups. The more unfavorable the attitude, from this point of view, the greater the social distance, and the more favorable the attitude, the less the social distance. Thus, the social distance between two intimate friends would be zero, and at the other extreme the attitude of a rabid anti-Semite toward Jews would represent maximum social distance." (23:164)

As applied to the field of teacher evaluation, professional distance refers to the frequency of disagreements and divergency between two professional workers on what constitutes the professional role of the good teacher. Professional distance may be illustrated simply as follows: Worker A believes a teacher should keep her pupils absolutely quiet during class time; Worker B believes pupils should be permitted considerable freedom. Such disagreements (and agreements) on the professional role of the good teacher constitute professional distance. As one increases the number of disagreements between any two professional workers, the professional distance increases. The greater the frequencies, the longer the professional distance.

The second aspect of the definition of professional distance suggests the measurement of the degree of divergency between the points of view of two professional workers. This aspect may be illustrated by Worker A, stating that she definitely believe a good teacher stands in front of the class when teaching; Worker B says she has no convictions on this issue, and Worker C says she definitely believes a good teacher stands in the rear of the room. Such a disagreement suggests that the professional distance is longer between Workers A and C than it is between A and B or B and C. The more divergent the points of view, the longer the professional distance; the less divergent the points of view, the shorter the professional distance.

Definition of the Professional Role of the Good Teacher.—The concept of the professional role of the good teacher also is an adaptation from the field of sociology. It is adapted from the term "social role" which Cuber defines as "the culturally defined patterns of behavior expected or required of persons in specific social positions.... behavior as used in this definition includes.... both overt acts and covert behaviors such as attitudes, values, and ideas." (10:232) Professional role may be similarly defined as the professionally determined behaviors expected or required of persons in a sepcific professional position, i.e., the position of classroom teacher. The role of teacher requires both covert and overt behaviors. In general, the covert behaviors may include desiring to teach, knowing subject matter, and believing in democracy. The overt behaviors are conducting lessons, manipulating teaching tools, and preparing reports.

When defining the behaviors required of persons playing the professional role of the good

*From the author's Ph.D. dissertation, University of Wisconsin, 1952; A. S. Barr, advisor.

TABLE I

SOME CORRELATIONS BETWEEN CRITERIA AND OTHER BEHAVIORS

Criteria	Other Behaviors	Correlations	Researcher
Pupil Change	Supervisory Rating	.36	Rolfe
	Practice Teaching Score	.13	Jones
	College Grades (4 yr. G. P. A.)	-.08	Jones
	Tests (American Council Psychological Examination)	-.10	Rolfe
Supervisory Rating	Practice Teaching Grades	.69	Bossing
	College Grades	.19	Bossing
	Tests (National Teachers Examination)	.51	Flanagan
Practice Teaching Grades	Supervisory Ratings	.69	Almy-Sorenson
	College Grades	.49	Almy-Sorenson
	Teaching Aptitude	.19	Seagoe
Pupil Ratings	College Grades	.03	Lins
	Tests (American Council Psychological Examination)	-.12	Lins

*All data for Table I is from A. S. Barr, "Measurement and Prediction of Teaching Efficiency: A Summary of Investigations," Journal of Experimental Education, XVI (June 1948).

teacher in contrast to the professional role of the poor teacher, one synthesizes all the best illustrations of "good" teaching he has seen or heard or read about. Teacher practices such as keeping the children absolutely quiet during class time, having them fold their hands while they listen to the teacher, and measuring the results of learning exclusively with standardized tests may be learned as "good" activities. On the other hand, dividing the class into small groups to work on individually chosen tasks, permitting as much freedom as possible, and measuring the results of learning by observing cooperative behavior patterns may be learned as "good" behaviors by a second professional worker who may be a rater of teachers. Likewise the personal traits that "good" teachers have, the modulated voice, the social poise, the ethical standards, all are learned by each professional worker to form his own concept of the "good" teacher and required of any person who would play the role of his "good" teacher.

How Concepts of the Professional Role of the Good Teacher Functions in Teacher Evaluation. — "Appraisal of any kind may be defined as an act of judgment, in which the judging implies both a criterion—a standard of some sort—and a pertinent description of what is being judged." (18:172) In the field of teacher evaluation the criterion is one's concept of the professional role of the good teacher or some aspect of it; the pertinent description is a concept of the teacher being judged, or some aspect of her performance in the role of teacher. When a teacher evaluates herself, she compares what she thinks she is to what she thinks she should be, i.e., her concept of the professional role of the good teacher. As a result of her comparison, she arrives at a qualitative and/or quantitative expression representing the distance between her two concepts. When evaluations are made by a person other than the teacher, the evaluator compares his concept of the teacher's performance with his concept of the professional role of his "good" teacher. His comparison also results in a qualitative and/or quantitative expression representing the distance between his two concepts. From this point of view, all measurement can be thought of as expressions of distance between the criterion and the concept of what is being evaluated. Concepts of the professional role of the good teacher function as the criterion in teacher evaluation.

How Concepts of Professional Distance Function in Teacher Evaluation—When professional distance—that is, disagreements between two professional workers on what constitutes the professional role of the good teacher—exists between the teacher and her evaluator, their evaluations are likely to be different because their criteria are different. Doing an excellent job of teaching in the eyes of the teacher is approximating her own concept of good teaching. If her concept of good teaching is decidedly different from the concept of good teaching held by her rater, i.e., the professional distance between them is long, the evaluation that the rater may give her is apt to be poor. If, on the other hand, her concept of good teaching is similar to that of her rater, i.e., the professional distance between them is short, the evaluation that the rater may give her is apt to be good. Thus it is hypothesized that the length of professional distance increases as teacher ratings decrease from good to average and from average to poor.

The Measurement of Professional Distance —Professional distance suggests the comparison of the concepts of the professional role of the good teacher as held by any two professional workers. To make such comparisons instruments were constructed to ascertain the overt and covert behaviors each professional worker expects from his "good" teacher. Specific teaching practices, teacher factors, and beliefs related to education were selected to appear on the instruments. Subjects were asked to respond by classifying each practice as good or poor; each factor as important or in significant; and each statement of belief as ones which they definitely believe or ones that they definitely do not believe. Step intervals were provided for indicating in between positions. Since distances between the teachers and their rater were sought, comparisons were made between the responses of the teacher and the responses of their raters. For each item on each instrument the distance between the two responses was assigned a weight value. To arrive at the total distance measured by the instrument, the weight values for all the items of that instrument were summed. *

Conditions Under Which the Hypothesis Will be Considered Substantiated—If the professional distance scores are lowest for the teachers rated good, and higher for the teachers rated average, and highest for the teachers rated poor, then the hypothesis will be considered to be substantiated, and professional distance as measured by these instruments may be considered as an indicator of professional ratings. If professional distance scores appear in some other pattern, the hypothesis will be considered as not supported by the evidence.

Summary—This study attempts to show the pattern of the lengths of professional distance as they exist between raters and the teachers they

*Instruments used in this study are described in detail in Section III. Copies of them appear as Appendices A through C which will be found in original thesis on file in the Library, University of Wisconsin, Madison, Wisconsin. Procedures for measuring professional distances are described in Section IV.

rate. Professional distance is defined as the number and divergency of the disagreements between the concepts held by two professional workers on what constitutes the professional role of the good teacher. Each worker learns from his own unique sequence of experiences his concept of the professional role of his "good" teacher. One's concept of the professional role of the good teacher is used as a criterion to evaluate one's own teaching and the teaching of others. When rating others, the resultant evaluations probably vary from good to poor as professional distances vary from short to long. Special instruments and procedures are used to measure professional distance.

SECTION II

The Method of Research

A MODIFIED form of the casual-comparative method of research was employed in this study. Two phenomena were investigated: one, a teacher considered a good teacher; the other, a teacher considered a poor teacher. The first modification recognized a middle group, called average, and believed to be between the two extremes.* Therefore, the absence of the first phenomenon was teachers rated average or poor; the absence of the second phenomenon was teachers rated average or good. The second modification assumed that circumstances attending the presence of the phenomena may exist in degrees, i.e., lengths of professional distance.

Some arbitrary limits were made for the study. All subjects were selected from elementary school faculties. Teachers in the group studied taught between grades one and six. Whether the relationship stated in the hypothesis exists among junior high and senior high school faculties is not a part of the study. Another limitation was made by definition. The professional role of the teacher was limited to a rather arbitrary set of teaching practices, teacher factors, and beliefs related to education. While each of the items seemed reasonable at the time of adoption, possibly other items and instruments for detecting other areas of disagreement between professional workers may be had.

Application of the Plan of Research. — The school systems of two communities were selected for the study. The first community with a population of approximately 100,000 was located on the coast of the Gulf of Mexico. When the study started in this community, there were 19 elementary schools for Anglo- and Latin American children. Two of the 19 schools were not accepted for the study: one had been established only a few weeks before the study was begun; the second was staffed by teachers who the principal felt could not be considered as either good or poor. Of the 17 schools selected, 15 were administered by their own principals. The other two were administered by one principal who felt competent to rate the teachers in both schools.

The second community had a population of approximately 70,000 and was located on the coast of Lake Michigan. Of the 14 elementary grade schools, 13 were selected for the study. One was not accepted because the teachers failed to cooperate. Among the 13 accepted, one elementary school was housed in a building that also housed classes for orthopedic and mentally handicapped children. The principal in this elementary school was administrator for all divisions in his building. Two elementary schools were housed in buildings along with junior high schools. The principals of the elementary schools were also principals of the junior high schools. Six small schools were administered by three principals, each principal serving as the head of two schools. Each of these three principals felt competent to rate the faculties in each of his schools. Each of the other schools was administered by its own principal. With the schools from the first community, the total group for study consisted of 30 elementary school faculties administered by 26 principals.

The principals of each of the 30 schools served as the raters of their teachers. Faculties of the schools ranged from 6 to 47 teachers. Each principal was asked to select from his staff(s) one of the best teachers, one of his average teachers, and one of his poor or ineffective teachers. These directions were employed so that he would use his own criteria in making his judgments. The selection of one good, one average, and one poor teacher was employed to secure a spread of his ratings. There is no claim that the teachers were selected on the same basis, nor that the teachers making up any one classification have anything in common other than their own principal's rating. With the selection of teachers, the subjects for the study consisted of 26 raters of teachers, 30 teachers rated good, 30 teachers rated average, and 30 teachers rated poor.

All subjects responded to the data gathering devices presented to them. On the first instrument the subjects classified controversial teach-

*Lamke (19), using a factor analysis technique, was led to believe that perhaps teachers considered average are not necessarily in between the two extremes.

er practices as "good", "poor", or "makes no difference". On the second instrument the subjects classified teacher factors on a five point scale from "of utmost importance" to "insignificant". On the last instrument the subjects indicated their pattern of beliefs related to education. The instruments are described in detail in Section III.

In each case the subject's cooperation was asked and received. None of the group was told the hypothesis being studied. Sufficient time was given to permit each person to respond at his leisure. Instructions on the mthod of response appeared on each measuring device. The qualifications of the persons and their responses suggested that the instruments were not misinterpreted. When omissions were considered oversights, the subjects were asked to complete the instrument. Omissions of one principal and three teachers, however, were due to a difference in point of view. In these cases it was inferred from their comments on the margin that they considered the items they omitted as "not making any difference". Their omitted responses were considered as such. These were few in number. One requirement specified that persons responding to the instruments would not converse about the study before or during the data collecting. Upon collecting the instruments from the schools, information on compliance with this request was asked. Responses indicated cooperation.

The study sought the relationship of professional distance to teachers' ratings. To determine professional distance, the responses of each rater were compared with those of the teachers he rated. Weight values were assigned to their disagreements. Three analyses of the assigned weights were made. Professional distance scores were computed by summing the assigned weights. Frequencies of disagreement scores were computed by counting the number of assigned weights. Item analyses were made by computing the professional distance and frequency of disagreement for each group of teachers for each item. Scores were compared to determine whether they substantiated the hypothesis.

SECTION III

Measuring Instruments

THREE DATA gathering devices were constructed and used in the study. They all sought to determine the subjects' concepts of the professional role of the good teacher, in terms of teacher factors, teaching practices, and beliefs related to education.*

The Evaluation of Teaching Practices.—The first of the three instruments dealt with teaching practices. Fifty-one of the practices appearing on the instrument were extracted from Table XLI, A Summary of Theory and Practice in Teaching Social Science, in A. S. Barr's Characteristic Differences in the Performance of Good and Poor Teachers of the Social Studies. (2:100f) The table includes data on the number of experts who consider the practice as good and also the number of experts who consider the practice as poor. In the construction of the instrument only those practices were selected on which the experts showed a marked degree of disagreement. Those practices on which the minority group of experts equalled ten or more percent of the majority group made up the first 51 items for the instrument. To this number of items were added the following two, which seemed to be controversial:

Measures results of learning by changed attitudes and behaviors; and
Measures results of learning by quality of pupils' projects and exercises.

Subjects were asked to categorize each of the total of 53 practices as "good", as "poor", or as "making no difference", i.e., neither good nor poor. The various methods of "scoring" the instrument are described in the following chapter, Analysis of Data.

The Evaluation of Teacher Factors.— An instrument to obtain the subjects' ranking of the importance of specific teacher factors was constructed by using the teacher factors that appear on the official rating scale used by the school system in the first community studied. The author of this study was also the author of the rating scale. Twenty-five teacher factors appearing on both devices are divided into four classifications: (1) the teacher as a person; (2) the teacher as a director of learning; (3) the teacher as a friend and counselor of students; and (4) the teacher as a member of a professional staff. Such items as "physically fit", "emotional stability" and "good speaking voice" appeared in the first classification. "Establishes attainable goals cooperatively with students", Has mastery of subject matter", and Skillful with a variety of tests and measurement devices" appeared in the second classification. "Builds a sense of security and personal worth in all students", and "Considers the development of the child as an individual more important than subject matter mastery" appeared in the third. "Guided by professional ethics" and "Actively cooperates in staff operations" appeared in the last classification. The subjects were asked to categorize each of the 25 factors into one of five categories: (1) utmost

*Copies of instruments will be found in Appendices A, B, C, in original thesis, Library of the University of Wisconsin.

major importance, (2) very important, (3) important, (4) usable, but not important, and (5) insignificant. It was assumed that those factors deemed important were those that the subject required of the person who plays the role of his good teacher. The "scoring" techniques used for this instrument are described in Section V.

The Inventory of Beliefs. — An instrument to determine which beliefs related to education were held by the subjects was constructed for this research. It consisted of 12 groups of statements of beliefs. Each group contained 10 statements. The names of the groups were Teacher-Pupil Relationships, Teaching Profession, Community Relationships, Objectives of Education, The Schools' Stand on Controversial Issues, Minority Groups, Democracy and Government, Economic Problems, Organized Labor, Religion, and Life Values. Each of the 120 statements began with the words, "I believe that...." Illustrations of the items are:

> I believe that the public schools have an obligation to provide sex education.
>
> I believe that the teachers who actively work for social and economic reforms are poorer teachers than those who stick to their own subject matter fields.
>
> I believe that another world war will come eventually, regardless of the steps we take to prevent it.
>
> I believe that it is un-American to peacefully advocate that the American Government should operate all steel, mining, transportation, and manufacturing industries.
>
> I believe that every person will set aside his principles when the rewards for doing so are high enough.

Subjects were asked to indicate on a special answer sheet their reactions to the statements as one of the following: (1) Yes, I definitely believe this statement; (2) I am inclined to believe this statement; (3) I cannot say whether I believe this statement or not because I have not made up my mind; (4) I am inclined not to believe this statement; and (5) No, I definitely do not believe this statement. On this instrument it was assumed that the subject's own belief was the one he required for the professional role of his good teacher.

The reliability of the inventory was studied by means of a test-re-test procedure. Thirty-four members of a class in Teacher Supervision responded to the inventory on two occasions with one week intervening. Analysis of their responses found that:

> The average number of the 120 items on which the class gave identical answers a week later was 73.8.
>
> The average number of the 120 items on which the class members reversed themselves was 15.8.

"Reversing themselves" was defined as indicating "Definitely believing" or "Inclined to believe" during one responding period and indicating "Definitely not believing" or "Inclined not to believe" the same item during the other responding period. Changes in response may be attributed to the effectiveness of the instructions during the intervening week, or to the degree of reliability of the instrument. Methods for "scoring" the instrument are described in a later Section of this report.

Summary. —The three data gathering devices used in this study have been described. They are (1) the Evaluation of Teaching Practices, (2) the Evaluation of Teacher Factors, and (3) the Inventory of Beliefs. All sought the subjects' concept of the professional role of his good teacher. Methods for "scoring" the instruments are described in a later Section of this report.

SECTION IV

Analysis of Data and Conclusions

AS HAS already been said, a modified form of the causal-comparative method of research was employed. It was hypothesized that the lengths of professional distances increase as teacher ratings decrease from good to average and from average to poor. Professional distance is suggested by the frequency and divergency of disagreements between the points of view on what constitutes the professional role of the good teacher. The greater the frequency of disagreements, the greater the professional distance; the greater the divergency of disagreements, the greater the professional distance.

Professional distance "scores" are computed for the number and divergency of disagreements between each rater and the teachers he rated. Scores are associated with the teachers, such as: The score of the teacher rated good is 145. Actually, the score is not the teacher's any more than it is the rater's, since it signifies the extent of the disagreements between them. However, for convenience, throughout this discussion, the rather lengthy expression, "the score for the distance between the rater and the teacher he rated good"; is abbreviated to "A's score". Likewise, "the score for the distance between the rater and the teacher he rated average" is "B's score", and "the score for the distance between the rater and the teacher he rated poor" is "C's score".

If the professional distance scores are low-

TABLE II

PROFESSIONAL DISTANCE SCORES FOR TEACHER PRACTICES

Code No. School	Teacher Rated Good	Teacher Rated Average	Teacher Rated Poor
1.	22	30	18
2.	28	17	24
3.	21	30	30
4.	28	36	36
5.	36	36	24
6.	25	26	27
7.	30	42	40
8.	30	35	36
9.	10	26	44
10.	47	30	51
11.	22	27	34
12.	33	39	32
13.	30	31	33
14.	16	25	19
15.	37	29	31
16.	25	29	32
17.	34	34	29
21.	24	32	42
22.	27	25	35
23.	34	29	26
24.	38	42	37
25.	25	34	37
26.	23	29	27
27.	39	24	27
28.	27	22	26
30.	33	30	34
31.	24	35	29
32.	27	18	29
33.	22	33	32
34.	26	28	42

*Code numbers were assigned to each school to assure their anonymity. It may be reported, however, that the numbers 1 through 17 represent the first community studied, and 21 through 34 represent the second community. The school in the second community that was to be designated number 29 was eliminated for reasons explained in Section II under the heading "Application of the Plan of Research".

est for the A teachers, higher for the B teachers, and highest for the C teachers, then the hypothesis will be considered as supported by the evidence. If the professional distance scores appear in some other pattern, then the hypothesis will be considered as not supported by the evidence.

This section is divided into three parts. Part One reports the analysis of data for professional distance (frequency and divergency of disagreements). Part Two reports an analysis for frequency of disagreements, without regard for their divergency. Part Three reports on item analysis for critical items. In each part the three instruments are analyzed separately.

Part One: Professional Distance

Teacher Practices. —The instrument for measuring professional distances for teacher practices consisted of 53 items which the subjects classified as "good", "poor", or "makes no difference", i.e., neither good nor poor. The responses of each teacher were compared with those of her rater. Differences in their responses were assigned the following weights:

Weight of 2: One professional worker classifying the practice as good; the other worker classifying it as poor.
Weight of 1: One professional worker classifying a practice as making no difference; the other classifying it as good or as poor.

Professional distance scores for teacher practices were computed by summing the assigned weights. Scores for A, B, and C teachers are shown in Table II.

Professional distance is shorter for A teachers than for either B or C teachers in 16 of the 30 schools, and longer for C teachers than for either A or B teachers in 13 of the 30 schools. In 20 schools it is shorter for A teachers than for C teachers.

Inventory of Beliefs. —The instrument to measure professional distance on beliefs in and related to education consisted of 120 statements to which the subjects selected one of the five following responses to be their answer: (1) Yes, I definitely believe this statement; (2) I am inclined to believe this statement; (3) I cannot say; (4) I am inclined not to believe this statement; or (5) No, I definitely do not believe this statement. The responses of each teacher were compared with those of her rater. The following weights were assigned to the differences between their responses to any one statement:

Weight of 4: One party definitely believing; the other, definitely not believing.
Weight of 3: One party definitely believing; the other, inclined not to believe.
Weight of 3: One party definitely not believing; the other, inclined to believe.
Weight of 2: One party inclined to believe; the other, inclined not to believe.
Weight of 2: One party responding that he cannot say; the other, either definitely believing, or definitely not believing.
Weight of 1: One party definitely believing; the other, inclined to believe.
Weight of 1: One party definitely not believing; the other, inclined to not believe.
Weight of 1: One party responding that he cannot say; the other, either inclined to believe, or inclined to not believe.

The assigned weights were summed to determine the professional distance score for beliefs. These are shown in Table III.

Professional distance for beliefs is shorter for A teachers than for either B or C teachers in 11 schools of the 30 studied. It is longer for C teachers than for either A or B teachers in 14 schools. In 19 schools, professional distance for A teachers is shorter than that for C teachers.

A second analysis of the belief inventory was made in which only assigned weights for differences that suggested opposition were summed. Such differences were those assigned weights of 2 in which one professional worker was inclined to believe and the other worker was inclined not to believe. The resultant scores, representing Oppositional Professional Distance for Beliefs, are shown in Table IV.

Oppositional professional distance for beliefs is shorter for A teachers than for either B or C teachers in 14 schools. It is longer for C teachers than for either A or B teachers in 17 schools. In 19 schools it is shorter for the A teachers than it is for the C teachers.

Teacher Factors. —The instrument to measure professional distance for teacher factors consisted of 25 items found on the teacher rating scale of the first community studied. Subjects were asked to rank the importance of each factor on the following five-point scale: (1) of utmost importance; (2) very important; (3) important; (4) usable, but not important; (5) insignificant. The responses of each teacher were compared with those of her rater, and the differences between each of their responses were assigned the following weights:

Weight of 4: One party ranking the factor of utmost importance; the other, as insignificant.
Weight of 3: One party ranking a factor as ut-

TABLE III

PROFESSIONAL DISTANCE SCORES FOR BELIEFS

Code No. School	Teacher Rated Good	Teacher Rated Average	Teacher Rated Poor
1.	171	195	210
2.	167	140	140
3.	148	129	168
4.	127	171	179
5.	157	179	201
6.	181	155	175
7.	176	158	138
8.	211	149	162
9.	180	172	203
10.	185	179	207
11.	146	173	186
12.	164	176	159
13.	143	133	142
14.	114	114	143
15.	140	197	176
16.	161	160	182
17.	182	174	165
21.	165	174	187
22.	190	176	194
23.	145	199	155
24.	229	215	193
25.	188	183	189
26.	157	179	159
27.	139	178	148
28.	102	174	193
30.	136	185	131
31.	179	179	161
32.	153	117	201
33.	165	181	176
34.	162	149	142

TABLE IV

OPPOSITIONAL DISTANCE SCORES FOR BELIEFS

School Code No.	Teacher Rated Good	Teacher Rated Average	Teacher Rated Poor
1.	99	133	162
2.	102	83	78
3.	93	60	115
4.	78	102	120
5.	75	84	109
6.	105	92	127
7.	130	124	88
8.	158	78	70
9.	101	79	111
10.	155	161	187
11.	92	101	119
12.	112	105	97
13.	104	86	92
14.	50	64	65
15.	104	156	138
16.	97	101	124
17.	126	136	156
21.	100	118	123
22.	158	99	150
23.	84	139	92
24.	203	129	147
25.	160	109	167
26.	113	156	90
27.	73	78	91
28.	53	124	139
30.	76	130	75
31.	118	113	78
32.	131	101	176
33.	103	125	133
34.	107	97	79

most importance; the other, usable but not important.

Weight of 3: One party ranking the factor very important; the other, insignificant.

Weight of 2: One party ranking the factor as important; the other, either of utmost importance, or insignificant.

Weight of 2: One party raning the factor as very important; the other, usable but not important.

Weight of 1: One party ranking the factor very important; the other, either of utmost importance, or important.

Weight of 1: One party ranking the factor as usable but not important; the other, either important, or insignificant.

The assigned weights for the differences between the responses were summed to arrive at a professional distance score for teacher factors. These scores are shown in Table V.

Professional distance for teacher factors is shorter for A teachers than for either B or C teachers in 13 of the schools studied. It is longer for C teachers than for either A or B teachers in 8 schools. In 16 schools, professional distance is shorter for A teachers than it is for C teachers.

A second type of analysis was made of the assigned weights. First, the assigned weights were marked plus (+) if the teacher ranked the factor as more important than did her rater. All other weights were marked minus (-). Second, the plus and minus weights were summed algebraically to arrive at a Compensated Score of Professional Distance for Teacher Factors. The assumption for such an analysis was that a teacher would not be rated lower if she thought a teacher factor less important than her rater did, provided that she thought some other factor more important than did her rater. Since plus and minus values were summed algebraically, Compensated Scores could be zero (0), a positive quantity, or a negative quantity. In interpreting such scores zero would suggest the absence of professional distance. Direction of professional distance would be indicated by the sign: positive scores suggest that the teacher classifies factors as more important than her rater; negative scores suggest that she classifies them as less important than her rater. Length of professional distance is indicated by the integer. In comparing two scores for length of professional distance and one is negative, only the integers are compared. Thus, in school number 34 the A teacher's score of plus 5 is considered to be a shorter professional distance than the C teacher's score of minus 7. Compensated Scores are shown in Tabel VI.

Professional distance, measured by such an analysis is shorter for A teachers than either B or C teachers in 12 of the schools studied. It is longer for C teachers than for either A or B teachers in 12 schools. In 18 schools the A teacher's professional distance is shorter than the C teacher's.

A third analysis was made of the assigned weights. Plus and minus signs were added similarly to the method applied in the second analysis, but then only the negative values were summed to arrive at a Less Than Score. The assumptions were that no compensation factor operated in any teacher being considered good, average, or poor; that her thinking a factor more important than her rater's opinion of the same factor has no bearing on her rating as a teacher; and that only her thinking factors to be less important than her rater's opinion of them bears on her rating as a teacher. Less Than Scores are shown in Table VII. In interpreting these scores, zero suggests the absence of professional distance, and the higher the score the greater the professional distance.

Such an analysis indicates that professional distance is shorter for the A teacher than for either the B or C teacher in 9 of the 30 schools studied. It is longer for C teachers than for A or B teachers in 11 schools. In 18 schools professional distance for A teachers is shorter than for C teachers.

Summary of Analyses for Professional Distance. —Professional Distance scores were computed according to a variety of described procedures for each of the three instruments. For each procedure (1) the number of schools in which the A teacher's professional distance score was lower than either the B or C teacher, (2) the number of schools in which the C teacher's professional distance score was higher than either the A or B teachers' score, and (3) the number of schools in which the A teacher's score was lower than the C teacher's score was found. These are summarized in Table VIII.

Conclusion. —The data do not completely support the hypothesis stated earlier. The shortest professional distance is not always between the rater and the teacher he rates good, nor is it always longest between the rater and the teacher he rates poor. Depending upon the method of analysis and the instrument, the number of schools in which the A teacher's score is lowest varies from 9 to 16. Similarly, the number of schools in which the C teacher's score is the highest varies from 8 to 17. The number of schools in which the A teacher's score is less than the C teacher's score varies between 16 and 20, depending on the method of analysis and the instrument.

Apparently, the behaviors required by the teachers' raters for persons performing the role of their good teachers function in a more com-

TABLE V

PROFESSIONAL DISTANCE SCORES FOR TEACHER FACTORS

School Code No.	Teacher Rated Good	Teacher Rated Average	Teacher Rated Poor
1.	19	35	26
2.	16	30	29
3.	9	19	14
4.	19	17	21
5.	25	31	22
6.	23	32	20
7.	22	29	17
8.	28	16	21
9.	28	19	19
10.	25	26	26
11.	21	30	40
12.	18	19	18
13.	19	15	17
14.	18	19	17
15.	14	17	13
16.	8	23	22
17.	12	21	16
21.	16	18	17
22.	15	18	11
23.	14	18	17
24.	18	25	11
25.	11	13	18
26.	12	13	17
27.	16	17	20
28.	16	15	22
30.	16	18	13
31.	16	18	25
32.	18	22	18
33.	26	31	20
34.	15	15	16

TABLE VI

COMPENSATED SCORES OF PROFESSIONAL DISTANCE FOR TEACHER FACTORS

School Code No.	Teacher Rated Good	Teacher Rated Average	Teacher Rated Poor
1.	9	33	22
2.	0	30	25
3.	- 1	- 9	- 4
4.	11	1	21
5.	25	31	18
6.	-11	-28	-18
7.	18	29	- 9
8.	28	- 6	19
9.	28	1	- 5
10.	21	26	- 4
11.	21	30	40
12.	-12	-11	-16
13.	17	15	- 1
14.	16	3	11
15.	- 8	- 6	- 1
16.	4	- 3	-16
17.	6	-15	- 1
21.	- 8	- 8	1
22.	3	- 6	9
23.	- 4	-10	- 5
24.	-16	-21	1
25.	- 1	3	10
26.	- 8	- 5	- 9
27.	0	2	4
28.	16	11	20
30.	2	12	- 7
31.	- 2	16	25
32.	4	6	8
33.	22	27	14
34.	5	5	- 7

TABLE VII

LESS THAN SCORES OF PROFESSIONAL DISTANCE FOR TEACHER FACTORS

School Code No.	Teacher Rated Good	Teacher Rated Average	Teacher Rated Poor
1.	5	1	2
2.	8	0	2
3.	5	14	9
4.	4	8	0
5.	0	0	2
6.	16	30	19
7.	2	0	13
8.	0	13	1
9.	0	9	12
10.	2	0	15
11.	0	0	0
12.	15	15	17
13.	1	0	9
14.	1	8	3
15.	11	12	7
16.	2	13	19
17.	3	18	8
21.	12	13	8
22.	6	12	1
23.	9	14	11
24.	17	23	6
25.	6	5	4
26.	10	9	13
27.	8	7	8
28.	0	2	1
30.	7	3	10
31.	9	1	0
32.	7	8	5
33.	2	2	3
34.	5	5	13

plex manner. Behaviors required by the raters may be classified into two or more classifications. One classification may be considered "core" behaviors, on which from the rater's point of view there is no controversy, and on which agreement is necessary for a teacher to be considered good by him. A second classification may be considered "peripheral" behaviors, on which there exists an unresolved controversy and on which differences in points of view may be understood and accepted. The instruments used in this study are loaded with items from the second classification, since controversial items were sought for them. It may be that instruments designed to require the rater to (1) state his position, and (2) state whether he would accept alternate behaviors, would possibly measure professional distance more precisely.

Another factor seems to be operating in the measurement of teachers. People do not disagree with one another with equal amounts of tact. It seems possible that a teacher who holds a point of view quite distant from that of her rater, may compensate for the possible conflict by being quite tactful about their differences. On the other hand, a teacher disagreeing on only a few issues may do so, so untactfully that a disproportionate weight is placed on her divergent points of view.

A third factor may help explain the findings. Apparently, some raters of teachers are more tolerant than others. Raters of teachers who are tolerant of conflicting points of view may accept frequent and divergent opinions and not permit them to affect their ratings. On the other hand, raters of teachers who are intolerant of conflicting points of view may base their ratings, in part, on teachers' non-acceptance of their points of view. These are only three suggestions that may clarify the findings.

Part Two: Frequency of Disagreements

Disagreements between raters and the teachers they rated were next analyzed without regard for the amounts of their divergencies. It was hypothesized that C teachers disagree with their raters more frequently than do B teachers, who in turn disagree with their raters more frequently than do A teachers. Many types of disagreements were found.

Procedures for analyzing each instrument are reported separately. The frequencies of each type of disagreement along with partial and complete totals of disagreements for each instrument are shown in Tables IX through XII. A summary of these tables, indicating the number of schools in which (1) the A teachers disagree less frequently than do either the B or C teachers, (2) the C teachers disagree more frequently than do either the A or B teachers, and (3) the A teachers disagree less frequently than do the C teachers, appears in Table XIII.

Teaching Practices. —Two types of differences were recognized in analyzing the responses to this instrument. When one professional worker classified the practice as "good" and the other worker classified it as "poor", the difference was considered an oppositional difference. All other disagreements were non-oppositional differences. Frequencies for oppositional, non-oppositional, and total differences for Teacher Practices for A, B, and C teachers are shown in Table IX.

Inventory of Beliefs. —Four types of differences were recognized in analyzing the responses to this instrument and are named and defined as:

Type 1: Non-oppositional differences
- a) One professional worker responding "cannot say, the other indicating any other response."
- b) One subject responding "Definitely believing", the other responding "Inclined to believe".
- c) One subject responding "Definitely not believing", the other responding "Inclined not to believe".

Type 2: Mild Opposition
One subject responding "Inclined to believe", the other responding "Inclined not to believe".

Type 3: Moderate Opposition
- a) One subject responding "Definitely believing", the other responding "Inclined not to believe".
- b) One subject responding "Definitely not believing", the other responding "Inclined to believe".

Type 4: Strong Opposition
One subject responding "Definitely believing", the other responding "Definitely not believing".

Frequencies for each type of disagreement for beliefs for A, B, and C teachers are shown in Table X.

Partial and complete totals or frequencies of disagreements for beliefs are shown in Table XI. Partial totals of types four and three are frequencies of strong and moderate disagreements. Partial totals of types four, three, and two disagreements are frequencies of oppositional differences. Total differences are the sums of all types of disagreements.

Teaching Factors. —In the analysis of this instrument for type of disagreements, the differences in the responses between the teachers

TABLE IX

FREQUENCIES OF TYPES OF DIFFERENCES ON TEACHER PRACTICES

School Code No.	Frequencies of Oppositional Differences			Non-oppositional Differences			Total Differences		
	A	B	C	A	B	C	A	B	C
1.	4	6	4	14	18	10	18	24	14
2.	2	5	4	24	7	14	26	12	19
3.	3	4	4	15	22	22	18	26	26
4.	5	8	7	18	20	22	23	28	29
5.	12	14	8	12	8	8	24	22	16
6.	4	5	5	17	16	17	21	21	22
7.	3	13	2	24	16	36	27	29	38
8.	6	8	9	18	19	18	24	27	27
9.	1	5	17	8	16	10	9	21	27
10.	15	9	19	17	12	13	32	21	32
11.	7	2	5	8	23	24	15	25	29
12.	11	13	10	11	13	12	22	26	22
13.	10	11	9	10	9	15	20	20	18
14.	4	7	5	8	11	9	12	18	14
15.	8	5	4	21	19	23	29	24	27
16.	5	8	4	15	13	22	20	21	26
17.	13	9	7	8	16	15	21	25	22
21.	5	5	9	14	22	24	19	27	33
22.	7	3	10	13	19	15	20	22	25
23.	8	5	5	18	19	16	23	24	21
24.	3	2	3	32	38	31	35	40	34
25.	4	5	8	17	24	21	21	29	29
26.	2	5	4	19	19	19	21	24	23
27.	13	8	7	13	8	13	26	16	20
28.	9	6	11	9	10	4	18	16	15
30.	11	11	8	11	10	18	22	18	26
31.	5	4	7	14	27	15	19	31	22
32.	6	3	10	15	12	9	21	15	19
33.	5	9	7	12	15	18	17	24	25
34.	9	5	11	8	18	20	17	23	31

TABLE X

FREQUENCIES OF TYPES OF DIFFERENCES FOR BELIEFS

School Code No.	Frequencies of Differences											
	Type 4			Type 3			Type 2			Type 1		
	A	B	C	A	B	C	A	B	C	A	B	C
1.	13	22	27	13	15	16	4	0	3	49	38	36
2.	14	5	6	12	15	14	5	9	6	50	47	47
3.	14	5	10	11	9	19	2	5	9	50	59	39
4.	1	7	11	12	24	20	19	1	8	41	54	54
5.	9	8	11	9	16	21	6	2	1	56	65	57
6.	5	6	10	25	14	23	5	13	9	60	42	41
7.	16	22	11	20	12	10	3	0	7	35	26	40
8.	33	10	11	8	12	8	1	1	1	31	49	57
9.	12	3	12	11	17	21	10	8	0	56	67	58
10.	25	35	37	15	7	13	5	0	0	26	16	20
11.	2	2	2	18	25	35	15	9	3	54	61	66
12.	15	9	14	16	19	13	2	6	1	37	56	48
13.	12	10	7	18	14	16	1	2	8	34	40	46
14.	3	0	4	6	10	11	10	17	8	57	47	51
15.	12	23	19	16	20	20	4	2	1	33	37	26
16.	8	7	9	21	21	24	1	5	8	58	55	51
17.	13	18	36	24	20	4	1	2	0	42	27	7
21.	5	14	13	22	18	21	7	4	4	47	40	43
22.	32	7	28	10	21	10	0	4	4	24	53	33
23.	8	19	7	16	21	16	2	0	8	51	49	54
24.	40	19	19	13	17	23	2	1	1	19	59	34
25.	32	15	32	10	13	13	1	55	0	22	55	16
26.	17	23	5	15	20	18	0	2	8	33	21	51
27.	8	7	4	9	14	15	7	4	15	50	75	47
28.	4	10	16	9	24	27	5	6	0	43	38	43
30.	6	9	3	16	30	13	2	2	12	57	49	51
31.	11	9	2	24	23	22	1	4	2	54	55	59
32.	18	17	27	19	11	22	1	0	1	19	17	22
33.	11	15	18	19	21	19	1	1	2	44	42	31
34.	8	9	4	23	13	19	3	11	3	47	45	49

TABLE XI

PARTIAL AND COMPLETE TOTALS OF TYPES OF DIFFERENCES FOR BELIEFS

School Code No.	Totals of Types 3 and 4			Totals of Types 2, 3 and 4			Total of All Types		
	A	B	C	A	B	C	A	B	C
1.	26	37	43	30	37	46	79	75	82
2.	26	20	20	31	29	26	81	76	73
3.	25	14	29	27	20	37	77	79	76
4.	13	31	31	32	32	39	73	86	93
5.	18	24	32	24	36	33	80	91	90
6.	30	20	33	35	33	42	95	75	83
7.	36	34	21	39	34	28	74	60	68
8.	41	22	19	42	23	20	73	72	77
9.	23	30	33	33	28	33	89	95	91
10.	40	42	50	45	42	50	71	58	70
11.	20	27	37	35	36	40	89	97	106
12.	31	28	27	33	34	28	70	90	76
13.	30	24	23	31	26	31	65	66	77
14.	9	10	15	19	27	23	76	74	84
15.	28	43	39	32	45	40	65	82	66
16.	29	28	33	30	33	41	88	88	92
17.	37	38	40	38	40	40	80	67	47
21.	27	32	34	34	36	38	81	76	81
22.	42	28	38	42	32	42	66	85	75
23.	24	40	23	26	40	31	77	89	85
24.	53	36	42	55	37	43	74	96	77
25.	42	28	45	43	33	45	65	88	61
26.	32	43	23	32	45	31	65	66	82
27.	17	21	19	24	25	34	74	100	81
28.	13	34	41	18	40	41	61	78	84
30.	22	39	16	24	41	28	81	90	79
31.	35	32	24	36	36	26	90	91	85
32.	37	28	49	38	28	50	57	45	72
33.	30	36	37	31	37	39	75	79	70
34.	31	22	23	34	33	26	81	78	75

and their raters were classified into two groups. The first group contained those differences in which the teacher thought the factor more important than did her rater; the second group contained those differences in which the teacher thought the factor less important than did her rater. The frequencies for both groups of disagreements together with the total number of disagreements for A, B, and C teachers are shown in Table XII.

Summary of Frequencies of Disagreements.—The hypothesis studied in Part II of this discussion was that C teachers disagree with their raters more frequently than do either A or B teachers, and that A teachers disagree with their raters less frequently than do either B or C teachers. Thirteen classifications of types of disagreements were recognized and studied. Tables IX through XII were analyzed to determine the number of schools in which (1) A teachers disagree with their raters less frequently than do either B or C teachers, (2) C teachers disagree with their raters more frequently than do either A or B teachers, and (3) A teachers disagree with their raters less frequently than do C teachers. These numbers of schools are shown in Table XIII.

Conclusions.—The data do not completely support the hypothesis stated earlier in Part Two of this analysis.

Depending upon the instrument and the classification of type of disagreements analyzed, the number of schools in which A teachers disagree less frequently than do either B or C teachers varies from 9 to 16. For 12 of the 13 classifications of types of disagreements the number of schools in which A teachers disagree less frequently was slightly less than 50 percent of the 30 schools studied. It appears that the number of times A teachers disagree with their raters more frequently than do either B or C teachers is slightly more than the number of times they disagree less frequently than either B or C teachers.

Depending upon the classification of types of disagreements studied, the number of schools in which C teachers disagree more frequently than do either A or B teachers varies from 8 to 14. This range would suggest that the number of schools is always slightly less than 50 percent of the schools studied. It appears, therefore, that the number of times C teachers disagree with their raters more frequently than do A or B teachers is slightly less than the number of times they disagree less frequently than do A or B teachers.

It is therefore concluded that the frequency of disagreement slightly increases as ratings increase from poor to average and from average to good.

Depending upon the classification of types of disagreements, the number of schools in which A teachers disagree with their raters less frequently than do C teachers varies from 13 to 19 (average 16). Such an average is slightly more than the 50 percent of the 30 schools studied. It is therefore concluded that the frequency of disagreement decreases slightly as ratings increase from poor to good. Such a conclusion does not entirely contradict the conclusion stated above which compared frequency of disagreement to ratings increasing from poor to average and from average to good. It is suggested that teachers rated average may not necessarily be between good and poor, a finding suggested by Lamke (19).

The interpretations of the findings of Part One, Professional Distance, seem equally applicable here. It was suggested that required behaviors for performing the role of the good teacher may exist as "core" behaviors, over which there may be no controversy and little or no disagreement, and "peripheral" behaviors, over which controversy and disagreement are acceptable. If this is so, then frequencies of disagreements over both core and peripheral required behaviors may have less meaning for comparisons with teacher ratings.

Secondly, the suggestion, made in the conclusions to Part One, that the manner of disagreement may be a potent factor along with the number of disagreements seems to apply to the findings on frequencies of disagreements. These are only two suggestions that may clarify the findings on frequency of disagreements.

Part Three: Item Analysis

The differences between the points of view of the teachers and their raters were analyzed by items. It was hypothesized that if A teachers disagreed less frequently and less divergently from their raters than did either B or C teachers on certain items, those items may be considered critical in that agreement with one's rater may be associated with a teacher being considered good by her rater. For convenience, such items were called "A teacher items". Similarly, it was hypothesized that if C teachers disagreed more frequently and more divergently from their raters than did either A or B teachers on certain items, those items may be considered critical in that disagreement with one's rater may be associated with a teacher being considered poor by her rater. Such items were called "C teacher items".

Teacher Practices.—This instrument consisted of 53 teacher practices which subjects classified as "good", "poor", or "makes no difference". The responses of the teachers were compared with those of their raters. Two types of disagreements between their responses

TABLE XII

FREQUENCIES OF TYPES OF DIFFERENCES FOR TEACHER FACTORS

School Code No.	Frequencies of Differences								
	Teachers thinking item more important			Teachers thinking item less important			Teachers thinking different from rater		
	A	B	C	A	B	C	A	B	C
1.	11	17	13	2	1	2	13	18	15
2.	8	18	18	5	0	2	13	18	20
3.	4	5	5	4	10	9	8	15	14
4.	13	8	17	4	6	0	17	14	17
5.	18	18	13	0	0	2	18	18	15
6.	5	2	1	12	17	13	17	19	14
7.	16	22	4	1	0	11	17	22	15
8.	17	7	14	0	10	1	17	17	15
9.	18	7	5	0	6	10	18	13	15
10.	21	22	9	2	0	13	23	22	22
11.	14	21	24	24	0	0	14	21	24
12.	3	4	1	12	11	12	15	15	13
13.	12	13	6	1	0	6	13	13	12
14.	13	9	10	0	7	3	13	16	13
15.	3	5	6	9	9	6	12	14	12
16.	5	9	3	2	12	13	7	21	15
17.	8	3	6	3	15	6	11	18	12
21.	3	5	8	10	11	7	13	16	15
22.	8	6	9	6	11	1	14	17	10
23.	4	3	4	8	11	9	12	14	13
24.	1	2	5	14	15	5	15	17	10
25.	5	7	11	6	5	4	11	12	15
26.	2	4	4	9	9	10	11	13	14
27.	7	8	9	7	6	7	14	14	16
28.	13	12	15	0	2	1	13	14	16
30.	7	13	3	6	2	9	13	15	12
31.	6	14	18	8	1	0	14	15	18
32.	18	18	11	2	1	3	20	19	14
33.	10	12	12	7	7	5	17	19	17
34.	10	9	6	5	4	10	15	13	16

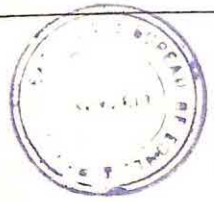


TABLE XIII

SUMMARY OF THE FREQUENCY TABLES

Instruments and Types of Differences	Number of Schools of the 30 Studied in which A disagrees less frequently than B or C	C disagrees more frequently than A or B	A disagrees less frequently than C
Teacher Practices			
Oppositional Differences	10	10	17
Non-oppositional Differences	13	10	18
Total Differences	16	11	18
Inventory of Beliefs			
Type 4 (Strong Opposition)	9	11	13
Type 3 (Moderate Opposition)	13	11	16
Type 2 (Mild Opposition)	8	9	13
Type 1 (Non-Opposition)	10	9	15
Sub-Totals			
Types 4 and 3	13	14	17
Types 4, 3, 2 (All Opposition)	13	14	19
Types 4 through 1 (All Differences)	13	10	18
Teacher Factors			
First Group	12	10	15
Second Group	9	10	17
Total Differences	12	8	14

were used in the item analysis. The first type were called oppositional disagreements and were defined as those in which one subject classified the practice as "good", while the other classified it as "poor". The second type of disagreements were called total disagreements and were defined as those in which the two professional workers responded differently from one another. For each item the number of oppositional disagreements and total disagreements between the A teachers and their raters were found. Similarly, for each item the number of oppositional disagreements and total disagreements for B and C teachers were found. These data are shown in Table XIV.

In analyzing Table XIV A teacher items were defined as those on which (1) the number of disagreeing A teachers is zero, while the number of disagreeing B or C teachers is one or more, or (2) the number of disagreeing A teachers is 50 percent or less of the number of disagreeing B teachers or C teachers, whichever is less. Items numbered 2, 3, 23, 24a, 24c, and 41 are A teacher items.

C teacher items were defined as those on which (1) the number of disagreeing C teachers is one or more, while the number of disagreeing A or B teachers is zero, or (2) the number of disagreeing C teachers is 200 percent or more of the number of disagreeing A or B teachers, whichever is larger. Items numbered 2, 4, 5, 6d, 19, and 22 are C teacher items.

Conclusions to Item Analysis of Teacher Practices. —The data supports the hypothesis that there are certain items on which A teachers disagree with their raters less frequently than do either B or C teachers. Six such items were found. They are:

2. Stands at the side of the room.
3. Stands at the rear of the room.
23. Organizes subject matter into psychologically arranged form (from pupils' experiences to logical generalizations).
24a. Assignments: page to page in textbook.
24c. Assignments: general topics and nothing more.
41. Measures results of learning by changed pupils' attitudes and behaviors.

Further, the data supports the hypothesis that on certain items C teachers disagree with their raters more frequently than do either A or B teachers. Six such items were found. They are:

2. Stands at the side of the room.
4. Sits at desk.
5. Sits on pupil's desk at front of room.
6d. Sits in pupil's seat at rear of room.
19. Provides for individual differences by differentiating assignments (the contract plan, unit instruction, level assignments, etc.).
22. Organizes subject matter in problem-project form.

One item, number 2, appears to be both an A and C teacher item. Apparently, agreement with one's rater on the evaluation of the practice of standing at the side of the room may be associated with a teacher being considered a good teacher, while disagreement with one's rater on the evaluation of this practice may be associated with a teacher being considered a poor teacher.

Individual items and their implications are discussed in more detail in the summary and conclusions to Part Three.

The Inventory of Beliefs. —This instrument consisted of 120 statements of beliefs related to education. Subjects were asked to select one of the following responses as their answer: (1) Yes, I definitely believe this statement; (2) I am inclined to believe this statement; (3) I cannot say; (4) I am inclined not to believe this statement; or (5) No, I definitely do not believe this statement. Responses of the teachers were compared with those of their raters, and weights were assigned to the differences as in Part One of this analysis.

Two approaches were followed in analyzing this instrument by items. The first approach considered frequencies of disagreements without regard to the degree of divergencies. The second approach considered both frequencies and divergencies. Four analyses were made of the data prepared for each of the two approaches. Critical items are quoted in the summary and conclusions of the item analysis of this instrument.

Analysis of Frequencies of Disagreements. —Two classifications of disagreements were used in the analysis. The first classification, called oppositional disagreements, was defined as those assigned weights of 4 or 3 plus those assigned a weight of two when one subject responded, "I am inclined to believe this statement," and the other responded, "I am inclined not to believe this statement." The second classification, called total disagreement, was defined as those on which the teacher responded differently from her rater. For each item the number of A teachers who opposed their raters and who responded differently from them were found. Similarly, the number of B and C teachers who opposed and responded differently from their raters were found for each item. These frequencies of A, B, and C teachers are shown in Table XV.

In analyzing the oppositional frequencies, A teacher items were defined as any item on which the number of A teachers who oppose their rat-

TABLE XIV

ITEM ANALYSIS OF TEACHING PRACTICES

Item Number*	Oppositional			Non-oppositional			Total		
	A	B	C	A	B	C	A	B	C
1.	3	2	2	14	16	13	17	18	15
2.	0	1	4	16	15	11	16	16	15
3.	1	2	2	17	15	14	18	17	16
4.	0	0	2	15	23	20	15	23	22
5.	0	0	1	17	14	13	17	14	14
6a.	0	1	0	16	15	21	16	16	21
6b.	2	3	1	15	18	23	17	21	24
6c.	3	2	2	12	16	19	15	18	21
6d.	2	0	4	14	16	17	16	16	21
7.	2	3	4	16	15	14	18	18	18
8.	0	0	0	13	12	14	13	12	14
9.	2	0	1	10	5	10	12	5	11
10.	3	5	3	17	19	16	20	24	19
11.	3	3	2	16	13	15	19	16	17
12.	4	5	4	18	15	18	22	20	22
13.	6	5	3	12	13	15	18	18	18
14	5	5	8	11	16	16	16	21	24
15.	7	5	8	11	10	13	18	15	21
16a.	9	5	8	6	10	9	15	15	17
16b.	8	3	3	1	2	4	9	5	7
16c.	9	10	13	3	4	4	12	14	17
16d.	5	6	5	6	4	5	11	10	10
16e.	4	2	7	7	8	5	11	10	12
17.	2	1	2	1	1	1	3	2	3
18.	7	7	7	6	6	3	13	13	10
19.	0	0	1	1	1	2	1	1	3
20.	5	5	5	5	7	3	10	12	8
21.	13	7	13	11	13	8	24	20	21
22.	1	0	4	8	7	8	9	7	12
23.	0	1	1	6	5	9	6	6	10
24a	2	3	5	2	5	4	4	8	9
24b.	6	6	6	3	9	7	9	15	13
24c.	1	3	3	4	5	3	5	8	6
24d.	9	5	6	4	11	5	13	16	11
24e.	0	1	0	3	3	5	3	4	5
25.	6	10	11	9	8	7	15	18	18
26.	3	5	5	4	7	4	7	12	9
27.	3	5	4	12	15	13	15	20	17
28.	3	5	5	5	6	7	8	11	12
29.	0	0	1	2	3	3	2	3	4
30.	0	0	0	6	6	6	6	6	6
31.	4	5	4	7	7	7	11	12	11
32.	7	8	7	10	12	11	17	20	18
33.	7	7	7	10	10	11	17	17	18
34.	5	4	5	8	11	11	13	15	16
35.	6	6	4	7	12	14	13	18	18
36.	3	2	3	6	6	6	9	8	9
37.	4	5	3	2	3	3	6	8	6
38.	3	6	4	3	3	5	6	9	9
39.	8	13	10	3	10	5	11	23	15
40.	6	3	6	4	12	8	10	15	14
41.	2	3	1	1	5	5	3	8	6
42.	6	8	3	6	8	8	12	16	11

*Three numbers (6, 16, and 24) of the sequence use letter suffixes, following the pattern of the source of the items. See beginning of Section III for a description of the instrument and the procedure followed in its construction.

Note: This table should be read as follows: Three A teachers opposed their raters on item one; two B teachers opposed their raters on item one; ... 14 A teachers differed but did not oppose their raters on item one; ... 17 A teachers differed from their raters for item one.

TABLE XV

ITEM ANALYSIS OF BELIEFS RELATED TO EDUCATION: FREQUENCIES OF OPPOSITIONAL AND TOTAL DIFFERENCES

Item Number	Oppositional Differences			Total Differences		
	A	B	C	A	B	C
1.	12	8	9	17	19	19
2.	9	12	11	20	21	23
3.	10	11	14	14	19	17
4.	11	13	10	22	22	20
5.	10	6	9	20	17	20
6.	7	7	5	20	15	14
7.	1	6	4	8	11	10
8.	9	12	11	15	21	17
9.	10	12	9	21	19	15
10.	0	0	0	9	13	9
11.	5	7	8	15	18	15
12.	10	8	8	20	20	16
13.	4	4	5	15	17	13
14.	4	4	5	9	15	9
15.	13	13	11	22	25	23
16.	13	14	15	25	28	24
17.	12	14	16	21	20	18
18.	1	4	2	10	12	10
19.	4	4	8	20	18	19
20.	9	9	9	20	22	23
21.	15	10	14	24	19	22
22.	8	10	9	22	25	21
23.	10	10	11	23	20	22
24.	12	11	19	21	22	27
25.	8	4	5	17	17	15
26.	9	12	17	22	21	23
27.	9	7	4	15	14	13
28.	3	3	3	8	16	9
29.	8	7	6	17	19	15
30.	12	10	11	18	23	21
31.	14	10	11	22	21	19
32.	12	11	11	23	23	21
33.	8	10	13	22	24	23
34.	13	10	8	18	20	18
35.	8	3	7	14	11	17
36.	16	13	13	25	24	22
37.	16	13	9	25	25	25
38.	2	4	0	14	15	20
39.	5	5	5	14	16	16
40.	5	10	8	17	22	18
41.	14	7	6	21	25	23
42.	2	4	1	20	18	18
43.	8	11	15	17	23	25
44.	4	6	9	12	12	15
45.	6	8	10	20	24	23
46.	7	5	8	17	21	23
47.	7	4	7	13	12	19
48.	7	5	7	23	22	20
49.	3	4	4	19	22	19
50.	7	7	8	16	18	19

TABLE XV (Continued)

Item Numer	Oppositional Differences A	B	C	Total Differences A	B	C
51.	8	11	9	22	25	23
52.	12	9	15	20	23	21
53.	3	6	1	15	25	18
54.	12	6	6	19	17	15
55.	9	9	9	21	23	24
56.	10	14	9	20	27	20
57.	10	7	13	20	20	23
58.	10	13	10	18	23	25
59.	11	9	14	25	26	25
60.	6	7	9	22	18	19
61.	9	7	10	21	23	22
62.	15	8	15	27	22	25
63.	8	10	8	20	24	23
64.	13	11	15	23	23	26
65.	9	12	12	25	24	22
66.	15	14	13	26	26	27
67.	5	5	5	17	17	18
68.	5	17	13	17	25	22
69.	12	13	15	24	25	24
70.	11	9	13	21	23	23
71.	14	14	14	22	24	24
72.	14	11	9	19	16	16
73.	9	7	7	25	21,	26
74.	6	7	6	20	20	20
75.	6	10	11	18	24	24
76.	8	8	7	18	19	17
77.	5	5	5	14	20	16
78.	5	6	5	15	21	19
79.	13	6	14	21	21	22
80.	11	11	12	24	25	21
81.	3	2	3	8	13	12
82.	10	17	7	23	24	25
83.	2	4	1	12	12	11
84.	1	4	2	16	20	12
85.	4	9	4	19	22	21
86.	12	11	9	24	23	19
87.	5	6	2	15	19	22
88.	11	8	8	20	15	14
89.	12	15	11	24	22	19
90.	13	9	15	20	23	23
91.	5	4	11	20	19	24
92.	9	7	10	25	18	22
93.	7	11	8	22	24	23
94.	9	11	11	26	22	23
95.	10	6	7	21	18	15
96.	7	6	5	17	16	18
97.	9	13	15	17	21	23
98.	1	2	1	11	11	13
99.	5	6	7	21	23	24
100.	8	9	13	22	22	24
101.	11	14	14	21	23	27
102.	11	12	11	23	21	23

TABLE XV (Continued)

Item Number	Oppositional Differences			Total Differences		
	A	B	C	A	B	C
103.	13	13	16	23	24	22
104.	13	15	17	23	22	22
105.	10	8	10	23	23	23
106.	13	14	13	24	23	22
107.	6	6	5	14	16	13
108.	16	12	11	24	20	20
109.	3	0	2	12	11	10
110.	12	12	15	24	24	24
111.	11	8	15	26	25	26
112.	5	7	11	15	19	19
113.	3	4	5	14	16	14
114.	4	2	3	12	13	12
115.	2	3	3	12	9	15
116.	7	6	9	21	20	21
117.	10	11	7	20	26	20
118.	2	5	3	8	13	12
119.	11	14	11	24	22	26
120.	1	0	3	11	12	13

ers was 67 or less percent of the number of B or number of C teachers, whichever is less, who oppose their rater. Seven items, those numbered 7, 18, 40, 44, 68, 75, and 84, fit the definition.

In analyzing the total differences frequencies, A teacher items were defined as any item on which the number of A teachers who disagree with their raters was 67 percent or less of the number of B or number of C teachers, whichever is less, who disagree with their raters. Items numbered 81, 105, 108, and 118 fit the definition.

In analyzing the oppositional frequencies for C teacher items, they were defined as any item on which the number of C teachers who oppose their rater is 150 or more percent of the number of A or number of B teachers, whichever is more, who oppose their raters. Four items, numbered 19, 24, 91, and 102, fit the definition.

In analyzing the total differences frequencies for C teacher items, they are defined as those on which the number of C teachers who disagree with their raters is 150 or more percent of the number of A or number of B teachers, whichever is more, who disagree with their rater. There are no such items.

Analysis of divergencies of disagreements.—Two analyses of the weights assigned to the disagreements were made. The first considered the weights assigned to oppositional disagreements. The second considered all assigned weights.

In the first analysis an oppositional weight for each item was computed for the A teachers by summing the weights assigned to the oppositional disagreements between these teachers and their raters. For the second analysis total weights for each item were computed for the A teachers by summing the weights assigned to all disagreements between these teachers and their raters. Similarly, oppositional weights and total weights were computed for each item for the B and C teachers. These data are shown in Table XVI.

In analyzing the oppositional weights, A teacher items were defined as those on which the oppositional weight of the A teachers was 67 or less percent of the oppositional weight of the B or C teachers, whichever is less. Nine items, numbered 7, 18, 40, 44, 68, 75, 84, 103, and 108, fit the definition. Only one item, number 103, was not detected using the analysis for frequencies.

In analyzing the total weights for A teacher items, they were defined as any item on which the total weight of the A teachers was 67 or less percent of the total weight of the B or C teachers, whichever is less. Items number 7 and 68, both found previously, were found to be critical.

In analyzing the oppositional weights for C teacher items, they were defined as those on which the oppositional weight for the C teacher was 150 or more percent of the oppositional weight of the A or B teacher, whichever is more. Items numbered 19, 24, 91, 112, and 120 were found to be critical. Items 112 and 120 had not been detected using other analyses.

In analyzing the total weights for C teacher items, they were defined similarly to the pattern above. No new items were found.

Conclusions to the Item Analysis of the Inventory of Beliefs.—The data seems to support the hypothesis that there are certain items on which A teachers disagree with their raters less frequently than do either B or C teachers. Twelve such items were found. They are:

7. I believe that pupils should be permitted to call teachers by their nicknames or given names.
18. I believe that teaching offers a wide variety of interesting experiences.
40. I believe that today's schooling makes too many students consider unskilled and semiskilled positions as not good enough for them.
44. I believe that teachers should teach students to side with the majority on controversial issues.
68. I believe that in hiring an individual for a job, it is often advisable to include race, color, and religion, in making your selection.
75. I believe that it is reasonable to fire a teacher who admits he is a Socialist.
81. I believe that private profits are essential to any successful economic system.
84. I believe that free enterprise has proved its superiority over other types of economic enterprises for America.
103. I believe that churches cause needless strife by over-emphasizing the differences among groups.
105. I believe that churches should take better care of their own parishioners rather than spend money on missions in foreign countries.
108. I believe that public schools should provide released time from classes for religious instruction.
118. I believe that working with people is better than working with things.

Further, the data seem to support the hypothesis that on certain items, C teachers disagree with their raters more frequently than do either A or B teachers. Six such items were found. They are:

TABLE XVI

ITEM ANALYSIS OF BELIEFS RELATED TO EDUCATION: WEIGHTED SCORES OF OPPOSITIONAL AND TOTAL DIFFERENCES

Item Number	Oppositional Weights			Total Weights		
	A	B	C	A	B	C
1.	38	27	12	44	41	42
2.	26	39	38	39	48	50
3.	37	34	45	43	44	48
4.	41	41	30	52	50	41
5.	36	21	32	47	32	44
6.	24	26	17	42	38	31
7.	4	20	15	12	25	21
8.	29	38	37	35	47	43
9.	36	37	29	50	44	37
10.	0	0	0	9	13	9
11.	18	26	24	29	38	32
12.	35	29	26	50	47	38
13.	14	11	18	28	28	26
14.	13	12	17	18	24	22
15.	44	43	36	54	58	49
16.	38	43	46	53	61	59
17.	40	44	49	50	47	51
18.	4	13	7	14	21	15
19.	13	13	24	35	31	39
20.	31	30	31	46	44	49
21.	53	30	47	65	42	58
22.	28	32	31	46	49	45
23.	35	30	38	49	41	51
24.	35	33	61	47	49	73
25.	26	13	15	37	28	28
26.	30	41	53	46	51	62
27.	27	24	12	33	31	22
28.	11	11	7	17	25	13
29.	23	23	20	32	36	30
30.	32	35	31	39	49	42
31.	48	32	37	57	48	47
32.	36	35	34	48	47	46
33.	26	32	44	45	49	58
34.	42	30	27	49	42	39
35.	27	8	25	34	17	36
36.	55	40	43	66	52	53
37.	45	37	28	57	54	48
38.	6	13	0	20	28	26
39.	16	17	17	27	31	29
40.	18	33	27	36	49	41
41.	46	24	19	56	50	43
42.	6	12	3	25	28	24
43.	27	38	49	37	50	62
44.	13	20	32	21	28	40
45.	23	27	31	41	48	47
46.	24	17	24	35	37	41
47.	25	12	28	33	31	42
48	23	16	28	46	40	46
49.	8	12	11	31	37	33
50.	22	25	30	33	39	47

TABLE XVI (Continued)

Item Number	Oppositional Weights A	B	C	Total Weights A	B	C
51.	24	36	30	41	52	48
52.	34	28	46	43	46	54
53.	9	22	2	26	44	26
54.	34	18	21	42	34	33
55.	25	28	27	41	45	44
56.	33	46	31	47	61	46
57.	32	21	43	46	39	56
58.	31	40	33	41	56	54
59.	35	27	45	53	47	60
60.	20	19	28	40	35	41
61.	31	23	35	47	43	54
62.	48	28	53	66	48	68
63.	27	30	28	45	50	47
64.	43	36	50	54	52	64
65.	28	38	37	49	53	49
66.	48	43	36	62	57	53
67.	18	19	17	36	33	35
68.	17	61	39	30	70	49
69.	37	37	43	32	50	53
70.	33	25	38	44	41	49
71.	46	44	45	54	54	56
72.	53	42	35	60	49	44
73.	26	24	26	48	44	57
74.	16	22	16	38	40	35
75.	21	34	34	36	52	55
76.	30	29	20	43	47	35
77.	16	16	15	31	35	30
78.	17	18	16	30	38	37
79.	43	20	48	52	39	59
80.	35	38	37	55	55	49
81.	11	6	11	17	20	22
82.	30	52	22	44	64	44
83.	6	16	4	17	25	16
84.	4	16	6	25	38	22
85.	12	29	13	32	49	35
86.	42	37	29	60	52	43
87.	17	22	7	30	39	35
88.	38	28	25	49	35	32
89.	40	45	36	53	53	45
90.	50	29	51	58	49	61
91.	16	13	36	33	32	52
92.	25	21	30	44	37	45
93.	19	34	26	38	48	43
94.	27	29	31	46	41	45
95.	30	20	22	46	37	31
96.	27	21	18	40	37	34
97.	33	44	44	42	52	53
98.	3	7	3	16	19	18
99.	17	20	22	36	42	42
100.	23	27	38	40	45	54

TABLE XVI (Continued)

Item Number	Oppositional Weights A	B	C	Total Weights A	B	C
101.	34	45	46	45	55	59
102.	36	38	36	49	49	48
103.	47	45	53	61	60	62
104.	43	46	57	56	54	62
105.	33	23	32	47	42	48
106.	48	50	48	62	64	58
107.	21	22	17	30	35	28
108.	51	39	36	61	49	45
109.	10	0	7	20	14	17
110.	40	43	53	55	59	63
111.	32	25	47	49	46	62
112.	15	20	39	26	34	47
113.	8	16	16	21	30	26
114.	14	7	11	26	22	24
115	7	9	10	19	17	25
116.	24	17	25	41	36	40
117.	33	39	22	45	59	41
118.	6	19	9	13	27	18
119	38	43	37	55	55	57
120.	3	0	10	14	14	22

19. I believe that in teaching, promotions are based on who you know rather than on what you know.
24. I believe that teachers should be free to use alcoholic beverages.
91. I believe that trade unions have done more harm than good in our industrial progress.
102. I believe that people who claim to be religious are less tolerant than people who do not claim to be religious.
112. I believe that most people will take advantage of you.
120. I believe that a large amount of money is a prerequisite to success.

A more detailed interpretation of the items and their implications is presented in the summary and conclusions to Part Three.

Teacher Factors. —This instrument consisted of 25 teacher factors which subjects classified on a five-point scale from "utmost importance" to "insignificant". The responses of the teachers were compared with those of their raters. When differences existed between the responses of the teacher and those of her rater, weights were assigned as in Part One of this section. Plus and minus designators were added similarly to the procedure used in the Compensated Score analysis.

Two approaches were followed in analyzing the positive and negative weights. The first approach considered frequency of weights without regard to divergencies. The second approach considered both frequencies and divergencies. Four specific procedures were followed to analyze the data prepared for each of the two approaches. Critical items are quoted in the summary and conclusions of the item analysis for Teacher Factors.

Procedures for Analyzing the Frequencies of Disagreements. —Following the first approach, frequencies of positive weights and negative weights were found for the A teachers for each item. Similarly, frequencies of positive and negative weights were found for the B and C teachers for each item. These data are shown in Table XVII.

In analyzing the frequencies of positive weights in Table XVII for A teacher items, it was hypothesized that A teachers would place more importance on the teacher factors than would B or C teachers. Therefore, for the first specific procedure, A teacher items were defined as items on which the number of A teachers who classified the factor as more important than did their raters is 200 or more percent of the number of B or number of C teachers, whichever is higher, who classified it as more important than did their raters. There are no such items.

For specific procedure number two, A teacher items were defined as items on which the number of A teachers who classified the factor as less important than did their raters is 50 or less percent of the number of B or number of C teachers, whichever is less, who classified it as less important than did their raters. Items 9, 12, and 16 fit the definition.

In analyzing Table XVII for C teacher items, it was hypothesized that C teachers would place less importance on the teacher factors than would A or B teachers. Therefore, for specific procedure number three, C teacher items were defined as items on which the number of C teachers who considered the factor more important than did their raters is 50 or less percent of the number of A teachers or number of B teachers, whichever is less, who consider the factor as more important than did their raters. There are no such items.

For specific procedure number four, C teacher items were defined as items on which the number of C teachers who classified the factor as less important than did their raters is 200 or more percent of the number of A or number of B teachers, whichever is more, who classified it as less important than did their raters. There are no such items.

Procedures for Analyzing Divergencies and Frequencies. —Following the second approach, a positive weighted score for A teachers was computed for each item by summing the positive weights assigned to the differences between the responses of the A teachers and their raters. Likewise, a negative weighted score for A teachers was computed for each item by summing the negative weights assigned to the differences between them. Similarly, sets of positive and negative weighted scores for each item were computed for the B and C teachers. These data are shown in Table XVIII.

Continuing the study of the same hypothesis proposed for the analysis of the frequencies of disagreements, for specific procedure number five, A teacher items were defined as items on which the positive weighted score of the A teachers is 200 or more percent of the positive weighted score of the B or C teachers, whichever is higher. No such item fits the definition. However, one item, number 21, seems to reverse the hypothesis to a marked degree in that positive weighted score for A teachers was 50 percent of the B teachers' score and 55 percent of the C teachers' score.

For specific procedure number six, A teacher items were defined as items on which the negative weighted score of the A teachers is 55 or less percent of the negative weighted score of the B or C teachers, whichever is less. Items numbered 9, 12, 24, and 25 fit the definition.

Continuing the C teacher aspect of the hy-

TABLE XVII

ITEM ANALYSIS OF TEACHER FACTORS: FREQUENCIES OF ASSIGNED POSITIVE AND NEGATIVE WEIGHTS

Item Number	A Teachers Pos.	A Teachers Neg.	B Teachers Pos.	B Teachers Neg.	C Teachers Pos.	C Teachers Neg.
1.	5	4	7	2	6	6
2.	18	5	20	3	16	6
3.	14	7	13	8	14	7
4.	13	6	12	9	12	9
5.	9	8	9	8	7	12
6.	12	5	13	7	9	7
7.	13	5	14	8	14	4
8.	11	7	10	14	10	8
9.	7	2	5	11	5	7
10.	14	6	18	9	13	8
11.	13	7	14	11	10	3
12.	17	1	22	4	17	3
13.	10	9	12	7	13	9
14.	8	8	12	10	10	10
15.	5	7	4	5	2	7
16.	11	2	12	5	12	4
17.	12	1	11	5	11	1
18.	14	6	8	10	11	7
19.	11	10	11	8	10	9
20.	14	4	13	8	11	6
21.	6	9	8	12	8	12
22.	7	10	15	5	8	10
23.	14	4	14	8	15	5
24.	16	5	14	7	12	9
25.	10	6	13	8	11	11

TABLE XVIII

TEACHER FACTORS: ITEM ANALYSIS OF WEIGHTED SCORES

Item Number	A Teachers Pos. Score	A Teachers Neg. Score	B Teachers Pos. Score	B Teachers Neg. Score	C Teachers Pos. Score	C Teachers Neg. Score
1.	5	6	8	3	7	7
2.	25	6	27	4	23	7
3.	17	7	17	9	18	9
4.	18	8	17	11	20	11
5.	13	8	11	9	9	15
6.	14	7	17	5	12	10
7.	18	7	17	12	19	4
8.	15	8	11	17	14	9
9.	8	2	6	11	6	8
10.	14	6	19	10	14	10
11.	14	9	18	14	15	5
12.	26	1	26	6	25	3
13.	13	9	18	8	17	10
14.	13	11	18	16	15	11
15.	6	7	5	8	3	7
16.	13	3	13	7	15	4
17.	13	1	12	5	11	1
18.	17	6	15	11	15	10
19.	14	12	15	11	13	13
20.	16	6	17	12	16	8
21.	7	15	14	18	13	15
22.	9	11	20	6	11	12
23.	16	4	19	8	19	6
24.	22	5	19	10	17	10
25.	14	6	20	11	16	16

pothesis stated above, for specific procedure number seven, C teacher items were defined as items on which the positive weighted score of the C teachers was 50 or more percent of the positive weighted score of the A or B teachers, whichever is lower. There are no such items.

For specific procedure number eight, C teacher items were defined as items on which the negative weighted score of the C teachers is 200 or more percent of the negative weighted score of the A or B teachers, whichever is higher. There are no such items.

Summary of the Eight Specific Procedures.—Differences between the points of view of the teachers and their raters were assigned positive weights when the teacher classified the factor as more important than did her rater, and negative weights when the teacher classified it as less important than did her rater. For each item of the Teacher Factor instrument frequencies of positive and negative weights were obtained for the A, B, and C teachers. Secondly, for each item positive weighted scores and negative weighted scores were obtained for the A, B, and C teachers. In greatly abbreviated form the eight specific procedures may be summarized with percentages and type of data indicated:

1. A is 200% of B or C using plus frequencies.
2. A is 50% of B or C using minus frequencies.
3. C is 50% of A or B using plus frequencies.
4. C is 200% of A or B using minus frequencies.
5. A is 200% of B or C using positive weighted scores.
6. A is 55% of B or C using negative weighted scores.
7. C is 60% of A or B using positive weighted scores.
8. C is 200% of A or B using negative weighted scores.

Conclusions to the Item Analysis of Teacher Factors.—The data seem to support some aspects of the hypothesis, stated earlier, and not to support other aspects of it. There are certain items on which the number of A teachers, who classify teacher factors as less important than did their raters, is decidedly less than the number of B or C teachers, who do so. Three such items were found. They are:

9. Provides for individual differences.
12. Has mastery of subject matter.
16. Obviously fair with pupils of all minority groups.

There are four items on which the degree of divergencies of the responses of the A teachers, who classified the Teacher Factors as less important than did their raters, is decidedly less than the degree of divergencies of the responses of the B and C teachers, who did so. The four items are 9 and 12, stated above, and:

24. Skillful in teacher-parent relationships.
25. Assists in care and improvement of the school equipment, buildings, and grounds.

There was one item on which the hypothesis seemed to be reversed. The degree of divergency of the A teachers, who classified the factor as more important than did their raters, was decidedly less than the degree of divergencies of the B or C teachers, who did so. This item is:

21. Offers thoughtful comments and criticisms for improvement of the school.

The data do not support the hypothesis in regard to C teacher items. On no items of the Teacher Factor instrument is the variation between the C teachers and their raters decidedly different from the variations between the A teachers and B teachers and their raters.

Individual items and their implications are discussed in more detail in the summary and conclusions to Part Three.

Summary and Conclusions to Part Three: Item Analysis.—Item analyses of all three measures used to study professional distances were made. It was thought that if on certain items the number of teachers rated good who disagreed with their raters was very low as compared with the number of teachers rated average or poor who disagreed with their raters, such items may be considered critical in that agreement on these items with one's rater may be associated with a teacher being considered good. Such items were called A teacher items. Similarly, it was thought that if on certain items the number of teachers rated poor who disagreed with their rater was very high as compared with the number of teachers rated good or average who disagreed with their raters, then those items may be considered critical in that disagreement on these items with one's rater may be associated with a teacher being considered poor. Such items were called C teacher items.

The disagreements between rater's and teachers' responses to each item were analyzed using several different approaches. For each approach definitions of A teacher and C teacher items were established and applied to the data. Of the 53 items on the Teacher Practices instrument, six were A teacher items, and six were C teacher items. Of the 120 items on the Inventory of Beliefs, twelve were A teacher items, and six were C teacher items. Of the

25 items on the Teacher Factor instrument, five were A teacher items, and none was a C teacher item. The number of critical items on the three instruments would suggest that of the 198 items studied, 23 are items on which teachers rated good agree with their raters to a decided degree more frequently than do teachers rated average or poor, and 12 are items on which teachers rated poor disagree with their raters to a decided degree more frequently than do teachers rated good or average. A total of 35 items were found to be critical.

The A teacher items, numbers 2 and 3 of the Teacher Practices instrument, suggest that teachers considered good agree with their raters on where or where not to stand while teaching. In contrast, four of the C teacher items found on the Teacher Practices instrument suggests that teachers rated poor disagree with their raters on where or where not to stand (item 2) and where or where not to sit (items 4, 5, and 6d).

The A teacher item, number 23 on the same instrument, suggests that teachers rated good tend to agree with their raters on the evaluation of the psychological arrangement of subject matter, while the C teacher item, number 22, suggests that teachers rated poor tend to disagree on the evaluation of organizing subject matter in problem-project form. Related to these practices seems to be the A teacher item, number 9 of the Teacher Factors instrument. Apparently, teachers rated good place more importance on the mastery of subject matter than do B or C teachers. Mastery of subject matter would be necessary for rearrangement of in into psychological unit form.

Agreement with one's rater on the role of the teacher concerning assignments seems to be a critical area. Item numbers 24a and 24c of Teacher Practices suggest that teachers rated good agree with their raters on the evaluation of page to page textbook assignments and general topic assignments with nothing more, while item 19 of the same instrument suggests that teachers rated poor disagree with their raters on the evaluation of providing for individual differences by differentiating assignments. The A teacher item number 12 of the Teacher Factor instrument seems related to this area. It suggests that A teachers tend to place more importance on providing for individual differences than do B or C teachers.

The last A teacher item of the Teacher Practices was measuring the results of learning by changed pupil attitudes and behaviors. Apparently, teachers rated good agree with their raters on the evaluation of this practice more frequently than do B or C teachers.

Agreement with one's rater on attitudes toward minority groups was found to be a critical area. Items 16 of Teacher Factors and 68 of the Inventory of Beliefs are in this category. Apparently, A teachers agree with their raters on the importance of being fair with pupils of all minority groups and believing or disbelieving that it is advisable to include race, color, and religion in hiring applicants for a position. The responses of C teachers did not differentiate themselves from B teachers in this area.

The topic of a teacher as a member of a professional staff offered three critical items. Assuming that differences among the variation between teachers' and their raters' classification of Teacher Factors indicate differences among A, B, and C teachers in their stress on their behaviors, it would seem that A teachers stress being skillful in teacher-parent relationships and assisting in the care and improvement of school equipment, building, and grounds more than do B or C teachers, and they stress offering thoughtful comments and criticisms for improvement of the school less than do B or C teachers.

Agreements and disagreements on religion was shown to be critical. In this area, three items, numbers 103, 105, and 108 of the Inventory of Beliefs, were A teacher items, and one item, number 102 of the instrument, was a C teacher item. Apparently, teachers rated good tend to agree with their raters more often and to greater extent on believing or disbelieving that churches cause needless strife by over emphasizing differences among groups, that churches should take greater care of their parishioners rather than foreign missions, and that schools should release school time for religious instruction. Teachers rated poor tend to disagree more frequently on believing or disbelieving that people who claim to be religious are more tolerant than people who do not claim to be religious.

A teachers tend to agree with their raters on items eulogizing the profession. Items number 18 and 118 of beliefs were found to be A teacher items. Apparently, teachers rated good agree with their raters on believing or disbelieving that teaching offers a wide variety of interesting experiences and that working with people is better than working with things. On the other hand, C teachers tend to disagree with their raters on beliefs suggesting less commendable attitudes toward the profession. Items 19 and 112 of beliefs were found as C teacher items. Apparently, teachers rated poor disagree with their raters on believing or disbelieving that in teaching, promotions are based on who you know rather than on what you know and that most people will take advantage of you.

Several critical items were found in an area that may be named Americanism and economics. On two items, numbers 75 and 81 on be-

liefs, teachers rated good were found to more frequently agree with their raters on believing or disbelieving that it is reasonable to fire a teacher who admits he is a Socialist and that private profits are essential to a successful economic system. Concerning labor, item number 40 was found as an A teacher item and number 91 and 120 of beliefs were found as C teacher items. Teachers rated good apparently tend to join their raters in believing or disbelieving that schools today make too many students consider unskilled and semi-skilled positions as not good enough for them, while teachers rated poor tend to disagree with their raters on believing or disbelieving that trade unions have done more harm than good in our industrial progress and that a large amount of money is a prerequisite to success. An A teacher item somewhat related to this area was number 44. Teachers rated good tend to agree with their raters in believing or disbelieving that teachers should teach students to side with the majority on controversial issues.

Two items dealt with the personal conduct of teachers. The A teacher item, number 7 of beliefs, suggests that teachers rated good tend to agree with their raters in believing or disbelieving that pupils should be permitted to call teachers by their given names. Item number 24 of beliefs suggests that teachers rated poor tend to disagree with their raters on believing or disbelieving that teachers should be free to use alcoholic beverages.

No attempt has been made in this study to isolate the teacher behaviors that are good or poor. It was the intention of the item analysis to determine on what items A teachers tend to agree with their rater more frequently than do B or C teachers, and on what items C teachers tend to disagree with their raters more frequently than do A or B teachers. Such A and C teacher items have been found and reviewed. It is not implied that the 35 items presented here are the only critical items for all raters. It is quite likely that certain raters have many more critical items while others have many less. Still others may have 35 different ones. The 35 critical items reviewed here were derived from 198 rather arbitrarily selected items by a rather arbitrarily, though logically, selected means.

It would seem that, in the evaluating of teaching, such minor aspects as where a teacher stands or sits may tend to attract the attention of the rater, especially if the teacher is sitting when he thinks she should be standing. However, it does not seem reasonable to believe that performing contrary to the rater's expectation on this issue would be sufficient reason to rate a teacher as poor.

Other items such as organization and mastery of subject matter, attention to individual differences, fairness to minority groups, and being a member of a professional group seem to be more significant items. Performing up to or surpassing the rater's expectations on such issues is undoubtedly sufficient reason to be considered good, while performing contrary to or short of expectations on these issues seems sufficient reason to be considered poor. If this is so, it is probably significant that teachers rated poor disagree with their raters more frequently than do A or B teachers on the evaluation of these teacher practices and on the importance of these teacher factors. It may be that pre-service and in-service education of teachers should place more emphasis on attitude formation toward these factors and practices along with knowledge and skill concerning them. That these factors are important has been suggested by other studies. That lack of performance in these areas are considered cause for failure and dismissal has been shown by Buellesfield (9), Madsen (21), and Nemec (22).

On the other hand, disagreements on religious items, eulogizing attitudes toward the profession, Americanism and economics, and personal conduct of teachers tend to remind one that performing the role of teacher is much more than performing instructional duties. It is probable that teacher evaluations are based, in part, on how closely one's behavior approximates the expectations of one's rater on certain critical areas. That proper behavior in these areas are important has been suggested by the studies of Edminston (12), Buellesfield (9), and Madsen (21). It is likely that pre-service education and orientational activities for beginning teachers may be modified to stress the broader aspects of the role of teacher. Further, it is possible that more compatible intra-staff relationships may be attained by selection and placement of personnel based on like beliefs on those issues found critical with the rater.

Lastly, it would seem that frequency and divergency of disagreements on certain issues between the rater and ratee are a factor in the teacher's evaluation. Therefore, to obtain a more accurate rating from the professional distance point of view, professional distances between rater and ratee should be measured and, if possible, kept constant, in the evaluation of teachers.

Section Summary. —Using a modification of the causal-comparative research method, the three data gathering instruments were analyzed from three points of view.

The first studied the hypothesis that professional distance, i.e., frequency and divergency of disagreements, increased as ratings decreased from good to average and average to poor. It was found that the data present here

did not completely support such a hypothesis. Depending upon the specific method of analysis and the instrument, the number of schools in which A teachers are a shorter professional distance from their raters than are the B or C teachers varied from 9 to 16 of the 30 schools studied. Likewise, the number of schools in which C teachers are further from their raters than are the A or B teachers varied from 8 to 17. It was suggested that raters may accept alternate points of view on items known to be controversial without decreasing the teacher's rating. Further, it was suggested that the manner of disagreeing may be an important factor along with the number of disagreements. Lastly, it was suggested that some raters are more tolerant of conflicting points of view than others.

The second approach studied the frequency of disagreements without regard to degree of divergencies. It was hypothesized that the teacher rated poor was the one who most frequently disagreed with the rater, while the teacher rated good most frequently agreed. The data seemed to indicate that frequency of disagreement slightly increases as ratings increase from poor to average and average to good. However, frequency of disagreement slightly decrease as ratings decrease from good directly to poor. Such conclusions do not entirely contradict one another. It is suggested that teachers rated average are not necessarily between good and poor, a hypothesis suggested by other research.

The third approach was an item analysis. It was hypothesized that on certain items A teachers agree with their raters more frequently and with less divergence than do B or C teachers, and that on certain items C teachers disagree with their raters more frequently and with greater divergence than do A or B teacher. Of the items studied, 23 were found to support the A teacher aspect of the hypothesis, and 12 were found to support the C teacher aspect. Of the 23 critical items associated with A teacher and rater agreement, six concerned Teacher Practices, 12 concerned beliefs related to education, and five concerned teacher factors. The twelve critical items associated with C teacher and rater disagreement were distributed equally between teacher practices and beliefs related to education. None was concerned with the importance of teacher factors.

It was suggested (1) that pre-service and in-service education of teachers increase the emphasis given to attitude formation toward such critical issues as individual differences, organization and mastery of subject matter, fairness to minority groups, and being a member of a professional staff, (2) that pre-service education and orientation of beginning teachers into the profession be modified to increase the emphasis on the broader aspects of teaching as a way of life, to include such areas as the teacher's role on religious issues affecting the school, eulogizing attitudes toward the profession, Americanism and economics, and personal conduct of teachers, (3) that, perhaps, more congenial intra-staff relationships may be attained by placement of personnel based on their points of view on critical issues, and (4) that professional distance on critical items must be considered in the evaluation of teachers to get more accurate ratings from the professional distance point of view.

SECTION IV

Summary and Conclusions

THIS STUDY attempts to show the relationship of professional distance to teacher ratings. Professional distance, adapted from the sociological term, social distance, is defined as the frequency and divergency between points of view held by professional workers on what constitutes the professional role of the good teacher. The greater the divergence and the more frequent the disagreements, the longer the professional distance; the lesser the divergence and fewer the disagreements, the shorter the professional distance. Professional role, adapted from the sociological term, social role, is defined as the overt and covert behaviors required of a person in a specific professional position. One's concept of the professional role of the good teacher, or aspects of it, serves as one's criterion to evaluate teaching. The evaluation of one's own teaching is an expression denoting the difference between one's concept of one's performance and one's concept of the professional role of the good teacher. The evaluation of another's teaching is an expression denoting the difference between one's concept of another's teaching and one's concept of the professional role of good teaching. If a teacher, who is approximating her own concept of teaching, is evaluated by a rater whose concept of good teaching is quite different from her's, the rating is apt to be poor. If the same teacher is evaluated by a rater whose concept of good teaching is similar to that of the teacher's, then the rating is apt to be good. Therefore, it is hypothesized that lengths of professional distance increase as ratings decrease from good to average and average to poor.

In reviewing researches in both education and sociology, the terms, professional distance and professional role of the good teacher, were not found. The concepts of social distance in place of "favorable" and "unfavorable" attitudes toward minority groups was reported by Bogardus in 1928. Assigning weights to de-

grees of social distance was reported by Dodds in 1935. In perhaps all educational research the presence of professional distance may be seen. The terms, traits, ratings, pupil change, test scores, and college grades, all imply specific behaviors. When these are used as criteria for good teaching, the behaviors required to attain desirable scores, ratings, pupil changes, etc., are the behaviors required of persons who would perform the professional role of the rater's concept of good teaching. Selected studies of the normative survey and correlational types were reviewed, and evidences of professional distances were pointed out. It was further suggested that correlations between sets of teacher ratings increase as the likelihood for professional distances decreases.

The methodology of research used in this study is a modification of the causal-comparative technique. The presence of the first phenomenon under investigation was a teacher being considered a good teacher, and the absence of this phenomenon was a teacher being considered average or poor. The second phenomenon was a teacher being considered a poor teacher, and its absence was a teacher being considered good or average. These definitions placed the average teacher in between the good and poor teachers. Frequencies and divergencies of disagreements with one's rater on the overt and covert behaviors required of a person who performs the role of the good teacher were studied as the circumstances attendant to the presence of both phenomenon.

In the application of the research technique, two communities were studied. From the first community 17 school faculties were selected. From the second 13 school faculties were selected. From each of the 30 faculties, the principal served as the rater of his teachers. He, in turn, selected a teacher he considered good, a teacher he considered average, and a teacher he considered poor. The groups of teachers were named A teacher, B teachers, and C teachers for convenience. All subjects were drawn from elementary school levels.

The measuring instruments for this study sought the points of view of the subjects on what overt and covert behaviors each required of the person playing the professional role of his good teacher. Three instruments were designed. The first consisted of 53 teacher practices which the subjects classified as "good", "poor", or "makes no difference", i.e., neither good nor poor. The second instrument was 120 statements of beliefs related to education. Subjects selected one of the following as their response: Yes, I definitely believe this statement; I am inclined to believe this statement; I cannot say; I am inclined not to believe this statement; and No, I definitely do not believe this statement. The third instrument consisted of 25 teacher factors which subjects evaluated on a five-point scale from "of utmost importance" to "insignificant". No claim was made that only these 198 items make up one's total concept of the professional role of the good teacher.

All raters and teachers responded to the instruments. The responses of the teachers were compared with those of their rater. Differences between the responses were quantified. Those differences suggesting little divergence were weighted a small value. Those suggesting a greater divergence, or opposition, were assigned a larger value.

Weights assigned to the differences were anlyzed from three points of view. The first considered professional distance, frequency and divergency of disagreements; the second considered frequency without regard to divergencies; the third was an item analysis. Professional distance scores were computed for the disagreements between each rater and each teacher he rated by summing the assigned weights. For convenience, the professional distance scores were associated with the teacher ratings. It was found that the data did not completely support the hypothesis, stated above. Depending upon the specific method of analysis and the instrument, the number of schools in which A teachers are a shorter professional distance from their raters than are the B or C teachers varies from 9 to 16 of the 30 schools studied. Likewise, the number of schools in which C teachers are a longer professional distance from their raters than are the A and B teachers varies from 8 to 17. It was suggested that raters may accept alternate points of view on items known to be controversial without decreasing the teacher's rating. Secondly, it was suggested that the manner of disagreeing may be an important factor along with the number of disagreements. Lastly, it was suggested that some raters of teachers are more tolerant of conflicting points of view than others.

The second analysis studied the frequencies of disagreements without regard to degree of divergence. It was hypothesized that A teachers disagree less frequently than do either B or C teachers, and C teachers disagree more frequently than do either A or B teachers. The data here presented seemed to indicate that frequency of disagreement slightly increases as ratings increased from poor to average and average to good. However, frequency of disagreement slightly decreases as ratings decreased from good directly to poor. Such conclusions do not necessarily contradict one another. It is suggested that teachers rated average are not necessarily between good and poor teachers.

The third analysis studied the 198 items on the three instruments. It was hypothesized that

on certain items A teachers agree with their raters more frequently and with less divergence than do either B or C teachers, and that C teachers disagree with their raters more frequently and with greater divergence than do either A or B teachers. Twenty-three items were found to support the A teacher aspects of the hypothesis and twelve items supported the C teacher aspect. It was suggested that (1) pre-service and in-service education of teachers emphasize attitude formation toward such critical items as individual differences, organization and mastery of subject matter, fairness toward minority groups, and being a member of a professional staff, (2) that pre-service and orientation activities for beginning teachers into the profession increase the emphasis on the broader aspects of teaching, (3) that, perhaps, more congenial intra-staff relationships may be attained by placing personnel on the basis of their points of view on critical items, and (4) that professional distance on critical items must be considered in the evaluation of teachers in order to get more accurate ratings from the professional distance point of view.

BIBLIOGRAPHY

1. Almy, H. C. and Sorenson, H. "A Teacher-Rating Scale of Determined Reliability and Validity," Educational Administration and Supervision, XVI (March 1930), pp. 179-86.
2. Barr, A. S. Characteristic Differences in the Teaching Performances of Good and Poor Teachers of the Social Studies (Bloomington, Ill.: Public School Publishing Co., 1929).
3. ———. "The Measurement and Prediction of Teaching Efficiency: A Summary of Investigations," Journal of Experimental Education, XVI (June 1948) pp. 204-84.
4. ———, and Emans, L. M. "What Qualities are Prerequisites to Success in Teaching?" Nation's Schools, VI (September 1930), pp. 60-4.
5. ———, and others. Supervision: Democratic Leadership in Improvement of Learning, second edition (New York: D. Appleton-Century Co., 1947).
6. Bogardus, E. S. Immigration and Race Attitudes (Boston: D. C. Heath and Co., 1928).
7. Bogardus, E. S. "The Measurement of Social Distance," in Readings in Social Psychology. T. M. Newcomb and E. L. Hartley, editors (New York: Henry Holt and Co., 1947).
8. Bousfield, W. A. "Students' Ratings of Qualities Considered Desirable in College Professors," School and Society, LI (February 1940), pp. 253-56.
9. Buellesfield, H. "Causes of Failure Among Teachers," Educational Administration and Supervision, I (September 1915), pp. 439-45.
10. Cuber, J. F. Sociology: A Synopsis of Principles (New York: D. Appleton-Century Co., 1947).
11. Dodd, S. C. "A Social Distance Test in the Near East," American Journal of Sociology, XLI (September 1935), pp. 194-204.
12. Edmiston, R. W. and Cahill, C. M. "What Does the Rural Community Expect of its Teachers?" Educational Administration and Supervision, XXVI (February 1940), pp. 98-102.
13. Encyclopedia of Educational Research, W. S. Monroe, Editor (New York: Macmillan Co., 1950). Revised Edition.
14. Encyclopedia of Social Sciences, E. R. Seligman, Editor (New York: Macmillan Co., 1934).
15. Good, C. V., and others. The Methodology of Educational Research (New York: D. Appleton-Century Co., 1941).
16. Greenwood, E. Experimental Sociology (New York: King's Crown Press, 1945).
17. Haggard, W. W. "Some Freshmen Describe the Desirable College Teacher," School and Society, LVIII (September 19-43), pp. 238-40.
18. Harris, C. W. "The Appraisal of a School —Problems for Study," Journal of Educational Research, XLI (November 1947), pp. 172-82.
19. Lamke, T. A. "Personality and Teaching Success," Journal of Experimental Education, XX (December 1951), pp. 217-57.
20. Lamson, E. F. "Some College Students Describe the Desirable College Teacher," School and Society, LVI (December 1942), pp. 6-15.
21. Madsen, I. N. "The Prediction of Teaching Success," Educational Administration and Supervision, XIII (January 1927), pp. 39-47.
22. Nemec, L. G. "Teacher Certification," Journal of Experimental Education, XV (September 1946), pp. 101-32.
23. Newcomb, T. M. Social Psychology (New York: Dryden Press, 1950).
24. Witty, P. A. "Evaluation of Studies of the Effective Teacher," in Improving Educational Research, official report of the American Educational Research Association (Washington, D. C.,: American Educational Research Association, 1948).
25. Wilson, L. and Kolb, W. Sociological Analysis (New York: Harcourt, Brace and Co., 1949).

26. Woodworth, R. S. and Marquis, D. G. Psychology (New York: Henry Holt and Co., 1947).

27. Young, P. V. Scientific Social Surveys and Research, second edition (New York: Prentice-Hall, 1949).

AN INVESTIGATION OF THE NEW YORK STATE REGENTS EXAMINATIONS IN SCIENCE

GEORGE GREISEN MALLINSON
Western Michigan College of Education
Kalamazoo, Michigan

JACQUELINE V. BUCK
Grosse Pointe Public Schools
Grosse Pointe, Michigan

Foreword

An investigation as extensive as the one reported herein obviously is not the work of one man. It represents the coordinated efforts of a number of science educators who have devoted hours of work and personal finances to accomplish a job that has long needed doing. Their rewards for all practical purposes are intangibles, chiefly the satisfactions from jobs well done. To give adequate credit to these workers is impossible. Suffice to say the list that follows contains the names of those to whom no adequate credit can ever be expressed verbally. Without their efforts, discussions about the New York State Regents Examinations in Science would fall purely into the realm of speculation and conjecture. The list follows:

Leo Alberti
Jacqueline V. Buck
Sidney V. DeBoer
Dale A. Fuelling
Lois M. Mallinson
David J. Miller
James L. Pellowe
John J. Schmitt
Fred J. Service
Wayne A. Stafford
Harold E. Sturm
Kenneth E. Summerer
Richard G. Telfer

Many other persons contributed time and effort in providing advice, criticisms and suggestions in various phases of the study. Among them are Mr. Wilton E. Baty, Chairman of the Committee on Regents Examinations, New York State Science Teachers Association; Mr. Hugh Templeton, Supervisor of Science Education of the University of the State of New York; Mr. Gordon E. Van Hooft, formerly President of the New York State Science Teachers Association; Dr. J. Cayce Morrison, formerly Coordinator of Research and Special Studies of the New York State Department of Education; Dr. Warren K. Findley, Director, Evaluation and Advisory Service, Educational Testing Service; Dr. Kenneth E. Anderson, Dean, School of Education of the University of Kansas; Dr. Francis D. Curtis, Professor-Emeritus of Education and of the Teaching of Science, University of Michigan; and Miss Agnes Hodahl, formerly New York State Representative of the National Science Teachers Association, Albany, New York.

SECTION I

THE EXPERIMENTAL DESIGN

The Problem

THE PROBLEM of this investigation is two-fold: (1) to investigate the attitudes of certain science teachers from the State of New York toward the New York State Regents Examinations in Science, and (2) to analyze and evaluate certain characteristics of the Regents Examinations for Biology, Chemistry, Earth Science, and Physics prepared for the examination periods of January 25, 1949; June 21, 1949; January 24, 1950; and June 20, 1950.

Background of the Study

On December 23, 1949, the director of this study met with Mr. Hugh Templeton, Supervisor of Science Education, of the University of the State of New York to discuss certain aspects of science teaching. During the course of the conversation the work of a committee of the New York State Science Teachers Association concerning the attitudes of science teachers of New York State toward the Regents Examinations was discussed. The committee, under the chairmanship of Mr. Wilton E. Baty, Huntington High School, Huntington, New York, had planned to poll a number of science teachers of New York State to obtain an objective analysis, for the first time, of their opinions of the Regents Examinations in Science.

Many factors, among them time, personnel, and finances, made it impossible for the committee to carry out its task. Hence, through the good offices of Mr. Templeton, the survey of opinions was delegated to the director of this study whose activities were still subject to the approval of Mr. Baty's committee and Mr.

Templeton.

Suffice to say the survey was duly carried out by Mr. David J. Miller of Lakeview Junior High School, Battle Creek, Michigan, and was prepared as a report for his master's thesis at the University of Michigan. With the approval of the University of the State of New York the report was subsequently published in an issue of Science Education. 1*

During the months that followed the initial stages of Miller's investigation, the director and Mr. Templeton met on a number of occasions to discuss the progress of the study. At one of these conferences, it was indicated that an investigation of the attitudes of the teachers was most desirable, but that it was unlikely that such an investigation would reveal the objective characteristics of the examinations. It was decided, therefore, that the director should prepare a research design that would provide a means for evaluating the objective characteristics usually evaluated in an examination as well as certain characteristics unique to the Regents Examinations in Science.

After several weeks the director submitted a design to Mr. Templeton. The design was studied by Mr. Templeton and other members of the State Department of Education who were likely to be concerned. After modifications were made in light of criticisms and suggestions, it was decided that a sampling of the Regents Examinations for Biology, Chemistry, Earth Science and Physics should be item-analyzed and the following factors considered:

1. A determination of the reliability, consistency and validity of the examinations.
2. A comparison between the scores obtained by students from small high schools and from large high schools.
3. A comparison between the scores obtained by girls and boys.
4. A determination of the levels of reading difficulty and vocabulary load of the various examinations.
5. An analysis of the types and frequencies of scoring errors made by the teachers who scored the examinations.
6. An analysis of the various test items on the examinations in order to determine their popularity, difficulty and discriminatory power.

The report that follows in Section II deals with the factors just mentioned.

*Footnotes will be found at the end of the article.

SECTION II

SAMPLING THE REGENTS EXAMINATIONS IN SCIENCE AND TALLYING THE SCORES ON THE ITEMS

The Problem

THE PROBLEM of this phase of the investigation is (1) to describe the procedure used in obtaining a representative sampling of Regent Examinations in Science for analysis, and (2) to describe the manner in which the scores on the examination items were talled and summarized.

Obtaining a Representative Sampling of the Regents Examinations in Science

In order to carry out the study it was necessary to obtain copies of the examination papers after they had been written by students in the State of New York. Copies of these papers were available since all the passing papers are forwarded to the State Department of Education and are stored for one year pending a possible review of the scores. Through the cooperation of Mr. Peter Muirhead, of the University of the State of New York, permission was received to obtain the needed examination papers. It was agreed that the number and type of papers needed would be sent to the investigators provided that students' names and locations were kept confidential, and that, at the completion of the study, all papers would be destroyed. (Suffice to say the agreement was scrupulously followed and all papers were subsequently destroyed.) One weakness in this phase of the study is obvious. Only passing papers were available and hence the study does not deal with any analysis of papers scored as failures.

The problem of sampling the vast number of papers was a difficult one since an accurate analysis of the parameter of the student population was impossible. A number of conferences were held with statisticians of the State Department of Education, University of Michigan and Western Michigan College of Education. Four assumptions were deemed defensible as a basis for making the sampling:

1. Complete bundles of papers turned in by schools should be selected instead of sampling individual papers. Thus considerations of cross-section of student population, and socio-econ-

omic level would be met beyond reasonable doubt.

2. The size of the bundles sent in to the State are influenced by the size and type of school, as well as geographical location. In order to take these factors into account in the sampling of papers, the distribution of sizes of bundles selected for analysis should be the same as the distribution of sizes of those turned in to the State.

3. The satisfaction of the first two assumptions would depend on the fact that at least 1500 papers for any one examination should be selected by a random selection of the proper sized bundles from the total group in storage.

4. In order to assure that the analysis would be representative for a field of science, papers from four consecutive examinations should be analyzed.

In the final study these considerations were met with the following modifications:

1. It was decided to analyze the examinations in the areas of Biology, Chemistry, Earth Science and Physics for January 25, 1949; June 21, 1949; January 24, 1950; and June 20, 1950. It will be noted that four consecutive examinations were not used. The examinations prepared for August were disregarded for two reasons:

   a) The persons taking Regents Examinations in Science in August are atypical of those taking them in January and June. Inordinate proportions consist of (1) poor students who failed the examination at one of the prior periods and are repeating the examination after further study in summer session, and (2) good students who are attempting to accelerate their high-school work via summer study.

   b) Frequently, 1500 examinations in each of the various areas of science are not forwarded to the State in August, and hence assumption 3 could not have been met.

2. It was decided to analyze approximately 2000 rather than 1500 papers for each of the examinations and periods in question in order to further insure a proper sampling. It was not possible to obtain 2000 examinations for Earth Science for any of the periods under consideration since in each case the total numbers of papers were less than 2000. However, all those turned in were analyzed

The process of sampling the papers was then undertaken through the joint efforts of the Examinations Bureau and the office of the Supervisor of Science Education. The papers were shipped at four separate periods, in each case one year after they had been forwarded to the State Department of Education. Table I lists the numbers of papers that were analyzed.

Tallying the Scores on the Regents Examination in Science

Any analysis of the examinations demanded, of course, a means for tabulating the points of credit obtained by the various students on the various items of the examinations. It was decided therefore to prepare a tally sheet that would be suitable for tallying the different examinations.

The Regents Examinations in Science are divided into two parts. Part I of all these examinations consists of fifty items of the modified true-false, completion and multiple-choice types. The following are examples:

A. Modified true-false type:

"30. Light is transmitted through a vacuum.
30. . . . . ."

(For each correct statement, the word true is written on the line following the item. If the statement is incorrect the term that must be substituted for the italicized term to make the statement correct is written on the line following the item.)

B. Completion type:

"5. An object with an excess of electrons is charged_____. 5 . . . . . ."

C. Multiple-Choice type:

"14. The liquid which contracts when heated from 3^{o} to 4^{o} C. is (1) alcohol (2) kerosene (3) mercury (4) water.
14 . . . . . ."

All the items on Part I give one point credit and are scored either right or wrong with no partial credit.

It was necessary therefore to develop a tally sheet on which the scores obtained by every student on every examination item could be tallied. Also later in the study it would be necessary to compute the coefficients of reliability of Part I of the examinations by means of the odds-evens technique. Hence it was decided to develop a tally sheet that would serve both these purposes. The section of the sheet for tallying Part I was divided so that the total number of points obtained on the even-numbered items could be computed separately from the total number obtained on the odd. Thus it was possible to obtain a score of twenty-five points on each of these halves. The numbers on the vertical ordinate of the tally sheet designate the number of the item; those on the horizontal ordinate, designate the various students whose papers were

tallied. Each student retained the same numerical designation throughout the tally sheet.

If an item on Part I were answered in error, a dot was placed in the proper square. When all items were so tallied, the total numbers of errors for the odd items, and for the even items, were totaled. From these totals were computed the total scores obtained by the students on the odd items, even items and on both. It should be stated here that papers of all students receiving the same total score were tallied on the same sheet. Thus papers scored (65) were tallied on one sheet, those scored 65 on another, those scored 66 on another, and so on. The scores obtained on the odd and the even items, together with the total score on Part I, were then entered in the appropriate spaces at the top of the tally sheet. A sample tally is shown on the next page.

The tally reads as follows:

For Part I, Student 1 gave incorrect responses to items 3, 21, and 37 for a score of 22 on the odd items. He gave incorrect responses to items 22 and 48 for a score of 23 on the even items. His total score for Part I is 45.

Part II, however, offered different problems with respect to tallying. It consists of eight or nine essay-type items, each worth ten points, from which the student may elect five for a possible total of fifty points. The items on Part II, however, vary from one another with respect to the numbers of parts they contain and the points of credit assigned to the parts.

In taking Part II of the examination, the student completely rejects either three or four items. It is also possible in some items to elect or to reject certain sub-parts. The former rejections are referred to in this study as "complete rejects" and the latter as "partial rejects." On the entire examination a total of one hundred points is possible if the student correctly answers all the items.

A second part on the tally sheet was developed to handle the various considerations in scoring Part II. The blocks on the second part of the tally sheet are numbered 1-8 to correspond with the numbers of the items on Part II.

In tallying Part II the numbers of the various sub-parts of the items were written in the left columns of the blocks. On this section, however, the actual scores of the students on the various parts of the items are entered in the appropriate squares. The first row across each block is labelled "Comp. Reject." A check is placed in this row if a student rejected an entire item. The spaces marked "Reject" are checked if a student rejected a subpart of an item (if given a choice). In the case of examinations having nine items, an additional block was affixed to the bottom of the sheet.

The total scores obtained by the students on the items they elected on Part II were then entered in the appropriate spaces at the top of the tally sheet. (See sample tally for Part I for space allotted to total score on Part II.)

A sample tally for Part II is shown on page 48.

The next task was to develop a means for summarizing the scores thus tallied so that they could be analyzed statistically. As a result three summary sheets were prepared.

Summary Sheet I was a duplicate of the top section of the large tally sheet except for spaces on the lower right in which were computed the horizontal totals. It was from this sheet that the scores were taken for computing coefficients of reliability for Part I, and for computing coefficients of consistency between the scores obtained by the students on Part I and Part II. A separate sheet was used for summarizing each tally sheet.

Summary Sheet II was designed to summarize the total number of errors, the total number of correct answers, and the average scores for each of the items on Part I. A separate sheet was used for all papers receiving the same total scores.

Summary Sheet III was designed to summarize the items on Part II. A separate sheet was used for all papers receiving the same total score. The following is an explanation of the meaning of the symbols in the blocks.

PP = total possible points. Computed by multiplying the value of the subpart by the number of persons who elected the sub-part.

PE = total points earned. This was the total number of points received by the persons electing the sub-part.

PM = total points missed. PP minus PE.

PR = the number of persons who rejected certain sub-parts of an item if a choice was given.

Av. Sc. = average score. This was obtained by dividing the total possible points (PP) by the number of persons who elected that sub-part of the item.

(Percentage score: This was entered in the margin and was obtained by dividing the points earned (PE) by the total possible points (PP).)

TA = total number of persons electing an entire item.

TR = total number of persons rejecting

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Odds	22																													
Evens	23																													
Part I	45																													
Part II																														

PART I ERROR TALLY

	1	2	3	4	5	6	7	8	9	10
1										
3	•									
5										
7										
9										
11										
13										
15										
17										
19										
21	•									
23										
25										
27										
29										
31										
33										
35										
37	•									
39										
41										
43										
45										
47										
49										
TOTAL WRONG	3									
SCORE	22									

	1	2	3	4	5	6	7	8	9	10
2										
4										
6										
8										
10										
12										
14										
16										
18										
20										
22	•									
24										
26										
28										
30										
32										
34										
36										
38										
40										
42										
44										
46										
48	•									
50										
TOTAL WRONG	2									
SCORE	23									

PART II

1	1	2	3	4	5	6	7	8	9	10
Comp. Reject										
A_1 Reject										
Points	1									
A_2 Reject										
Points	2									
B_1 Reject										
Points	2									
B_2 Reject										
Points	2									
B_3 Reject										
Points	2									
Reject										
Points										
Reject										
Points										
Reject										
Points										
Total	9									

2	1	2	3	4	5	6	7	8	9	10
Comp. Reject	X									
A_1 Reject										
Points										
A_2 Reject										
Points										
B_1 Reject										
Points										
B_2 Reject										
Points										
B_3 Reject										
Points										
Reject										
Points										
Reject										
Points										
Reject										
Points										
Total										

3	1	2	3	4	5	6	7	8	9	10
Comp. Reject										
A Reject										
Points	2									
B Reject										
Points	1									
C Reject	.									
Points										
D Reject										
Points	2									
E Reject										
Points	2									
F Reject										
Points	1									
Reject										
Points										
Reject										
Points										
Total	8									

The tally for Part II reads as follows: Student 1 elected item 1 having sub-parts A_1, A_2, B_1, B_2, and B_3. He received 1 point on part A_1, and 2 points on each of the other parts. He decided to reject completely item 2 having parts A_1, A_2, B_1, B_2, and B_3. He elected item 3 having parts A, B, C, D, E and F and chose to reject, as was his privilege, part C. He received one point of credit on Parts B and F, and 2 on each of Parts A, D and E for a total of 8 points.

an entire item.

TA plus TR equals the total number of persons whose papers were summarized on the sheet.

The nine blocks, of course, were designed for summarizing separately the scores obtained on the various sub-parts of the eight or nine items found on Part II of the examinations. In the left columns of the various blocks were entered the numbers and letters designating the various sub-parts of the items.

The uses made of the computations found on the various Summary Sheets will be indicated in later pages of this report.

Sample Summary Sheets that are filled out are shown on the next three pages. (Note: A check mark below a grade denotes an error in the scoring of a student's paper.)

SECTION III

THE RELIABILITY, CONSISTENCY AND VALIDITY OF THE REGENTS EXAMINATION IN SCIENCE

The Problem

THE PROBLEM of this phase of the investigation is (1) to describe the methods used in computing the reliability, consistency and validity of the Regents Examinations in Science, and (2) to report the results of these computations.

Methods Employed

It was obvious that any measure of the reliability, consistency and validity of the Regents Examinations in Science would involve computing coefficients of correlation. The device chosen for use in this study was the Pearson r. The scatter diagram technique cited by Guilford[2] was employed for making the computations.

As will be described more fully later, a small number of classes were used, in certain computations, for grouping the data into intervals. Thus the estimates of correlation were lowered to some degree. It was decided therefore to correct for errors in grouping, using data prepared by Peters and Van Voorhis.[3] In terms of a formula the correction is as follows:

$$r_c = \frac{r}{C_x C_y}$$

in which r_c is the corrected coefficient of correlation, r is the coefficient of correlation as computed from the coarsely grouped data, and C_x and C_y are the correction factors based on the number of class intervals in X and Y respectively. The use of this correction seems justified since the assumptions underlying its use were met.

Computing Coefficients of Reliability

There are three common methods for computing coefficients of reliability, namely, the split-half, alternate-form, and multiple-administration. Since only one form of a Regents Examination in Science is prepared, and that administered only once, the only method suitable for use in this study was the split-half.

This method, however, could not be used in determining the reliability of an entire Regents Examination in Science. A casual survey of a sample examination indicates clearly that the split-half method is applicable to Part I only. Hence, it was decided to compute the reliability of Part I without regard for the scores on Part II.

The scores from Summary Sheets I were then transferred to a scatter diagram. Those the students obtained for the odd items on Part I were tallied on the horizontal ordinate, those for the even, on the vertical. The tallies were made on the basis of two point intervals, namely 6-7, 8-9, and so on up through 25 points. The point score of 6 was set as the lower limit since it was highly improbable that a lower score on either the odd or even items on Part I would have appeared on a passing examination paper.

The coefficients of reliability were then computed and adjusted with the Spearman-Brown formula, and corrections were made for coarse grouping of data. Table II lists the results.

Computing Coefficients of Consistency

Since it was not possible to use the scores on Part II for computing coefficients of reliability, it was decided to compute coefficients of correlation between the total scores the students obtained on Part II and those they obtained on Part I. It was assumed that such computations (referred to here as "coefficients of consistency") might show the relationship between the abilities of students to answer correctly the "factual" items on Part I and their abilities to answer the "thought" items on Part II.

The coefficients of consistency were computed and corrected for coarse grouping of data. Table III lists the results.

Conclusions

Insofar as the techniques used in this phase of the investigation may be defensible, the fol-

SUMMARY SHEET I

Biology REGENTS EXAMINATION FOR June 20, 1950

Summary of Scores of Individuals on 60 Tests Scored at 78

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Odds	19	21	20	23	24	21	22	22	20	21	18	22	20	21	21	21	21	24	19	23	20	24	20	21	21	19	22	18	20	19	18	20	20	20	20
Evens	21	18	18	13	16	19	19	17	21	19	21	20	21	24	20	22	22	22	19	20	20	21	17	19	24	17	19	18	23	21	17	17	18	20	19
Part I	40	39	38	41	40	40	41	39	41	40	39	42	41	45	41	43	43	46	38	43	40	45	37	40	45	36	41	36	43	40	35	37	38	40	39
Part II	38	39	40	37	38	38	38	39	37	37½	39	36	37	33	37	35	35	32	40	35	38	33	41	38	33	42	37	42	35	38	43	41	40	38	39

	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70
Odds	20	21	19	19	19	24	20	20	22	19	16	20	20	20	17	22	20	16	23	19	20	19	20	21	22										
Evens	20	17	20	18	20	21	21	17	18	19	20	19	19	19	19	18	20	18	22	22	18	19	23	23	20										
Part I	40	38	39	37	39	45	41	37	40	38	36	39	39	39	36	40	40	34	45	41	38	38	43	44	42										
Part II	38	40	39	41	39	33	37	41	40	40	42	39	39	39	42	38	35	44	33	37	40	40	35	35	36										

	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	TOTALS
Odds																															1223
Evens																															1177
Part I																															2400
Part II																															2280½ (2283½)
																															4680½

SUMMARY SHEET II

Biology REGENTS EXAMINATION FOR June 20, 1950

Summary of Part I of 60 Tests Scored at 78

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Errors	2	2	1	14	0	22	10	18	1	17	2	17	10	12	9	4	6	1	9	21	9	11	9	8	43
Correct	58	58	59	46	60	38	50	42	59	43	58	43	50	48	51	56	54	59	51	39	51	49	51	52	17
Average Score	.97	.97	.98	.77	1.00	.63	.83	.70	.98	.72	.97	.72	.83	.80	.85	.93	.90	.98	.85	.65	.85	.82	.85	.87	.28

Number	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Errors	3	28	2	23	2	18	14	2	1	10	20	12	16	1	4	34	25	9	21	12	24	10	31	7	13
Correct	57	32	58	37	58	42	46	58	59	50	40	48	44	59	56	26	35	51	39	48	36	50	29	53	47
Average Score	.95	.53	.97	.67	.97	.70	.77	.97	.98	.86	.67	.80	.73	.98	.93	.43	.58	.85	.65	.80	.60	.83	.48	.88	.78

SUMMARY SHEET III

Biology REGENTS EXAMINATION FOR June 20, 1950

Summary of Part II of 60 Tests Scored at 78

%		PP	PE	PM	PR	Av.Sc.
.60	1a	90	72	18		4.80
.85	1b	60	51	9		3.40
	Sum	150	123	27		8.20
	TA	15	TR	45		

%		PP	PE	PM	PR	Av.Sc.
.67	2a	204	137	67		2.69
.95	2b	102	97	5		1.90
.73	2c	102	74	28		1.45
.96	2d	102	98	4		1.92
	Sum	510	406	104		7.96
	TA	51	TR	9		

%		PP	PE	PM	PR	Av.Sc.
.83	3a	156	130	26		3.33
.81	3b	78	63	15		1.62
.86	3c	78	67	11		1.72
.76	3d	78	59	19		1.51
	Sum	390	319	71		8.18
	TA	39	TR	21		

%		PP	PE	PM	PR	Av.Sc.
.83	4a	78	64½	13½	1	1.65
.80	4b	74	59	15	3	1.59
.89	4c	76	67½	8½	2	1.78
.75	4d	66	49½	16½	7	1.50
.71	4e	50	35½	14½	15	1.42
.83	4f	56	46½	9½	12	1.66
			(322½)			
	Sum	400	321½	77½	40	9.60
	TA	40	TR	20		

%		PP	PE	PM	PR	Av.Sc.
.61	5a	66	40	26		1.82
.93	5b	44	41	3		1.86
.76	5c	110	84	26		3.82
	Sum	220	165	55		7.50
	TA	22	TR	38		

%		PP	PE	PM	PR	Av.Sc.
.83	6a	76	63	13		1.66
.75	6b	114	85	29		2.24
.66	6c	152	100½	51½		2.64
.66	6d	38	25	13		.66
	Sum	380	273½	106½		7.20
	TA	38	TR	22		

%		PP	PE	PM	PR	Av.Sc.
.78	7a	104	80½	23½		3.10
.31	7b	78	24	54		.92
.89	7c	78	70	8		2.69
	Sum	260	174½	65½		6.71
	TA	26	TR	34		

%		PP	PE	PM	PR	Av.Sc.
.85	8a	74	63	11		1.70
.87	8b	74	64	10		1.73
.68	8c	37	25	12		.68
.59	8d	185	110	75		2.97
			(262)			
	Sum	370	263	108		7.08
	TA	37	TR	23		

%		PP	PE	PM	PR	Av.Sc.
.88	9a	50	39	11	7	1.56
.89	9b	60	47	13	2	1.57
.68	9c	56	38	18	4	1.36
.82	9d	56	46	10	4	1.64
.75	9e	60	45	15	2	1.50
.61	9f	38	23	15	13	1.21
	Sum	320	238	82	32	8.84
	TA	32	TR	28		

TABLE I

NUMBERS OF EXAMINATIONS RECEIVED

Field of Science	Dates of Examinations				Total
	Jan. 1949	June 1949	Jan. 1950	June 1950	
Biology	1984	2095	2441	2092	8612
Chemistry	1960	1974	2160	1920	8014
Earth Science	1624	1899	1419	1645	6587
Physics	1682	2142	2278	2002	8104
Total	7250	8110	8298	7659	31317

TABLE II

COEFFICIENTS OF RELIABILITY FOR PART I OF THE REGENTS EXAMINATIONS IN SCIENCE

Date of Examination	Biology	Chemistry	Earth Science	Physics
January 1949	.78 ± .02	.76 ± .02	.70 ± .02	.75 ± .02
June 1949	.64 ± .02	.77 ± .02	.57 ± .03	.77 ± .02
January 1950	.77 ± .02	.74 ± .02	.39 ± .04	.79 ± .02
June 1950	.58 ± .02	.82 ± .02	.85 ± .01	.77 ± .02

TABLE III

COEFFICIENTS OF CONSISTENCY BETWEEN PARTS I AND II OF THE REGENTS EXAMINATIONS IN SCIENCE

Date of Examination	Biology	Chemistry	Earth Science	Physics
January 1949	.61 ± .01	.72 ± .01	.45 ± .01	.65 ± .01
June 1949	.55 ± .01	.60 ± .01	.48 ± .01	.82 ± .01
January 1950	.49 ± .01	.62 ± .01	.40 ± .02	.58 ± .01
June 1950	.49 ± .01	.65 ± .01	.42 ± .02	.50 ± .01

lowing conclusions seem valid:

1. Most of the coefficients of reliability found in Table II (ten of sixteen) were .75 or higher. These values are considerably higher than similar computations for teacher-made tests, at least insofar as the available research evidence indicates. The values, however, are somewhat lower than those usually found for coefficients of reliability for standardized achievement tests in science.
2. It must be kept in mind that the coefficients of reliability were computed for only the fifty points on Part I of the various tests rather than for the total one hundred points. Had some technique been available for including the points obtained on Part II, a higher degree of reliability might have been indicated.
3. The Regents Examinations in Science are prepared for three examination periods during the year, whereas a standardized examination remains essentially the same from year to year except for occasional revisions. Thus any criticisms of the reliability must be tempered in the light of this fact.
4. The coefficients of consistency in Table III for total ranges of scores were not generally as high as the coefficients of reliability found in Table II. Thus there is no assurance that Part I and Part II have the same relative degree of difficulty for the students.

Computing Coefficients of Validity

Ordinarily the validity of a measuring instrument is determined by comparing the scores obtained on that instrument with criterion data for the factor being measured. In the case of this study only one measure for each student was available. Thus there was no possibility of comparing the scores the students obtained on the Regents Examinations in Science with the scores they obtained on any other measure. Hence, it was decided to use an internal criterion.

The various examinations were submitted to at least five members of the National Association for Research in Science Teaching, who taught science education at the college or university level and who, at one time or another had worked on some phase of test construction. They were asked to identify on Part II for each of the examinations, an item or part of an item that could be considered defensibly to be a measure of each of the following objectives:

a. Ability to apply scientific principles
b. Possession of scientific attitudes
c. Ability to employ problem-solving skills (elements of scientific method)

If a majority of the specialists identified a certain item or part of an item, as being a measure of one of the objectives listed above, the scores on these items or parts thereof, were considered tentatively as being criterion data for these objectives. The items thus identified were then resubmitted to all the specialists who were asked to examine carefully the items (or parts) and to make their judgments as to whether they could be considered defensibly as being measures of the objectives. Only those items considered suitable by four of the five judges were retained. In some cases the specialists failed to identify a suitable criterion item. In the case of these examinations computations for validity were omitted.

Table IV lists the various items or parts of items that were identified as being suitable for the intended purpose. The scores obtained by the students on the total test were then plotted on the horizontal ordinate of the correlation chart and those obtained by the students on the criterion items were plotted on the vertical ordinate. The computations for validity were then made and corrected in the same manner as those for consistency. Tables V, VI, and VII list the coefficients of validity thus computed.

Conclusions

Insofar as the techniques used in this phase of the investigation may be defensible, the following conclusions seem valid:

1. An examination of the Table VI indicates that for none of the examinations for Chemistry and Earth Science was an item or part of an item on the respective Part II considered to be a measure of the possession of scientific attitudes.
2. Tables V through VII indicate that the coefficients of validity for the various objectives differ greatly. In some cases they can be considered high, in other cases, low. Thus it is difficult to generalize with respect to the validity of the different Regents Examinations in Science.
3. In general, one might state that the Regents Examinations in Science are better measures of the ability to apply scientific principles, than to use elements of scientific method. The data for scientific attitudes are not sufficiently extensive to warrant a conclusion.
4. As compared with the validity of teacher-made tests that of the Regents Examinations in Science is, in general, high. As compared with certain standardized tests the validity seems to rate rather favorably, although in some cases it seems low.
5. It should be kept in mind that the methods for computing the coefficients of validity are not

TABLE IV

NUMBERS OF ITEMS, OR PARTS OF ITEMS, USED AS CRITERION DATA*

Examination and Date	Number of Item or Part of Item		
	Understanding of Scientific Principles	Possession of Scientific Attitudes	Ability to Use Elements of Scientific Method
Biology			
Jan. 1949	8 b, c	---	7 c
June 1949	9	5 a_1, a_2	2 a
Jan. 1950	3	8	9
June 1950	4	5 c	1
Chemistry			
Jan. 1949	8	---	3
June 1949	4	---	2
Jan. 1950	4	---	---
June 1950	6	---	4
Earth Science			
Jan. 1949	---	---	8
June 1949	3	---	8
Jan. 1950	5	---	---
June 1950	7	---	4
Physics			
Jan. 1949	8	---	2 a
June 1949	8	6	2 a_1, a_2
Jan. 1950	8	7	2
June 1950	6	---	8

*The dashes indicate that no item was judged as being suitable to serve as criterion data.

TABLE V

COEFFICIENTS OF VALIDITY (UNDERSTANDING OF SCIENTIFIC PRINCIPLES)

Date	Biology	Chemistry	Earth Science	Physics
January 1949	.56 ± .01	.73 ± .01	No Adequate Criterion	.55 ± .03
June 1949	.72 ± .01	.67 ± .01	.20 ± .02	.40 ± .02
January 1950	.20 ± .02	.55 ± .01	.34 ± .02	.54 ± .02
June 1950	.58 ± .01	.70 ± .01	.30 ± .02	.63 ± .01

TABLE VI

COEFFICIENTS OF VALIDITY (SCIENTIFIC ATTITUDES)

Date	Biology	Chemistry	Earth Science	Physics
January 1949	No Adequate Criterion	No Adequate Criterion	No Adequate Criterion	No Adequate Criterion
June 1949	.57 ± .02	No Adequate Criterion	No Adequate Criterion	.55 ± .01
January 1950	.50 ± .03	No Adequate Criterion	No Adequate Criterion	.58 ± .02
June 1950	.37 ± .02	No Adequate Criterion	No Adequate Criterion	No Adequate Criterion

TABLE VII

COEFFICIENTS OF VALIDITY (ELEMENTS OF SCIENTIFIC METHOD)

Date	Biology	Chemistry	Earth Science	Physics
January 1949	.41 ± .01	.49 ± .01	.15 ± .02	.47 ± .02
June 1949	.46 ± .02	.45 ± .01	.48 ± .01	.40 ± .02
January 1950	.58 ± .02	No Adequate Criterion	No Adequate Criterion	.61 ± .02
June 1950	.65 ± .02	.61 ± .01	.23 ± .01	.49 ± .02

the ones ordinarily used. Hence, the above conclusions must be evaluated in terms of this fact.

SECTION IV

AN INVESTIGATION OF THE RELATIVE ACHIEVEMENTS OF MALES AND FEMALES AND OF RURAL AND URBAN STUDENTS, ON THE REGENTS EXAMINATIONS IN SCIENCE[4]

The Problem

THE PROBLEM of this phase of the investigation is to determine whether achievement on the Regents Examinations in Science (1) varies with the sex of the student, and (2) varies with the size of the school in which the student is enrolled.

Methods Employed

In practically every level of science education two questions arise frequently. The first deals with the relative achievements of boys and girls, and the second, with the relative achievements of rural and urban students. The vast numbers of examination papers that were tallied in this investigation made possible a study of the questions.

The first step in this phase of the investigation was to tabulate, for each separate bundle of papers, the following information:

1. The field of science for which the examination was prepared
2. The date for which the examination was prepared
3. The name and location of the school from which the examination papers were received
4. The population of the school for the school year in which the examination was prepared. (Note: It was decided to accept the school enrollment for the year 1948-49 as the base population, since it did not, in most cases, differ sufficiently from that of 1949-50 to make a distinction.)

The information for the first three points was found on the examination papers. That for point four was found in the Forty-Sixth Annual Report on the Education Department for the School Year ending June 30, 1949, Volume 2, entitled Statistics (Albany: University of the State of New York, 1950, pp. 353). The information for the populations of high schools in the City of New York was not listed in this publication. It was, however, obtained from the Supervisor of Science Education of the State Department of Education.

The first step was to tabulate separately for males and females the scores made on Part I, Part II and on the total examination. This task was relatively simple.

However, it was more difficult to classify a school as being urban or rural in character. Hence, rather than to tally scores according to this method of classification it was decided to tally them according to the size of the school in which the students were enrolled. For the purpose of this investigation the Southern Michigan classification for size of high school was used. As set up in the Handbook of the Michigan High School Athletic Association for the School Year of 1952-53 the classification is as follows:

1. Class A - over 800 students
2. Class B - 325 to 799 students
3. Class C - 150 to 324 students
4. Class D - less than 150 students

This classification, however, did not prove to be completely satisfactory. It did not seem reasonable to classify the students of a school with an enrollment of 1000 with those from a large New York City high school, such as James Madison School with a population of over 6000. Therefore an additional classification was added, namely AA, or schools with a population of over 1500.

A copy of the sheet on which the scores were tabulated is shown on the next page.

In order to determine the significance of the variances that might exist among the scores on the basis of sex and class of school, it was decided to use the analysis of variance technique with the double entry table described by Lindquist.[5] However, there was a great inequality in replications since papers were sampled from those contributed by the various classes of schools in the same proportions as the various classes contributed papers to the total number sent into the state. As a result the factor of non-orthogonality was present in the design. The procedure for the final "F" test was the one suggested by Snedecor[6] for use with two-way classifications with unequal replications in which corrections are made for non-orthogonality.

A copy of the analysis sheet used for this purpose is shown on page 59.

Table VIII presents the results of the computations just described.

It may be noted that in several cases the variance with respect to interaction are significant. These occur on Parts I and II, and total score of the Biology Examination for June 1950; on Parts I and II of the Chemistry Examination for January 1949; on Part II of the Chemistry Examination for June 1949; and on Parts I and II, and

NAME OF EXAMINATION _____ DATE OF EXAMINATION _____

NAME OF SCHOOL:

_____

_____

POPULATION:

1948-49 _____

1949-50 _____

MALE				FEMALE			
Student	Part I	Part II	Total	Student	Part I	Part II	Total
1				1			
2				2			
3				3			
4				4			
5				5			
6				6			
7				7			
8				8			
9				9			
10				10			
11				11			
12				12			
13				13			
14				14			
15				15			
16				16			
17				17			
18				18			
19				19			
20				20			

ANALYSIS OF VARIANCE (Two-way classification)

Examination _____ Date _____ Part(s) _____

Class of School	Boys	Girls	$T_{(row)}$	$T_{(row)}^{2}/N$
AA	ΣX^2 = $\frac{T^2}{N}$ =	ΣX^2 = $\frac{T^2}{N}$ =		
A	ΣX^2 = $\frac{T^2}{N}$ =	ΣX^2 = $\frac{T^2}{N}$ =		
B	ΣX^2 = $\frac{T^2}{N}$ =	ΣX^2 = $\frac{T^2}{N}$ =		
C	ΣX^2 = $\frac{T^2}{N}$ =	ΣX^2 = $\frac{T^2}{N}$ =		
	ΣX^2 = $\frac{T^2}{N}$ =	ΣX^2 = $\frac{T^2}{N}$ =		
$T_{(col)}$			T =	
$T^2_{(col/N)}$				

A - $\Sigma\Sigma\Sigma X^2$ = _____

B - $\Sigma\Sigma T^2/N$ = _____

C - $T^2_{(row)}/N$ = _____

D - $T^2_{(col)}/N$ = _____

E = T^2/N (Correction) =

A - E = _____ (SS_T)

B - E = _____ (SS_{cells})

C - E = _____ (SS_R)

D - E = _____ (SS_C)

Summary Table

Source of variance	df	Preliminary sum of squares	Corrected sum of squares	Mean square	F	Tab F 5%	Tab F 1%
Sex							
Class							
Interaction							
Sub-total							
Within							
Total							

TABLE VIII

ANALYSES OF VARIANCES

Examination	Part(s)	Source of Variance	F	Interpretation		
				Significance	High	Low
Biology, January, 1949	I	sex	4.96	sig.	boys	girls
		class	3.39	sig.	AA	B
		interaction	.40	not sig.	--	--
	II	sex	2.05	not sig.	--	--
		class	5.44	very sig.	AA	D
		interaction	1.34	not sig.	--	--
	I & II	sex	4.03	sig.	boys	girls
		class	6.35	very sig.	AA	D
		interaction	.57	not sig.	--	--
Biology, June 1949	I	sex	.07	not sig.	--	--
		class	12.7	very sig.	AA	A
		interaction	1.34	not sig.	--	--
	II	sex	4.23	sig.	girls	boys
		class	10.84	very sig.	AA	A
		interaction	.49	not sig.	--	--
	I & II	sex	1.66	not sig.	--	--
		class	15.3	very sig.	AA	A
		interaction	.47	not sig.	--	--
Biology, January 1950	I	sex	9,96	very sig.	girls	boys
		class	21.5	very sig.	AA	D
		interaction	1.0	not sig.	--	--
	II	sex	.38	not sig.	--	--
		class	13.3	very sig.	AA	D
		interaction	.47	not sig.	--	--
	I & II	sex	3.39	not sig.	--	--
		class	21.2	very sig.	AA	D
		interaction	.38	not sig.	--	--

TABLE VIII (Continued)

Examination	Part(s)	Source of Variance	F	Interpretation		
				Significance	High	Low
Biology, June 1950	I	sex	7.10	sig.	boys	girls
		class	47.2	very sig.	AA	A
		interaction	7.31	very sig.	--	--
	II	sex	4.6	sig.	boys	girls
		class	11.3	very sig.	AA	D
		interaction	4.19	sig.	--	--
	I & II	sex	6.7	sig.	boys	girls
		class	24.5	very sig.	AA	D
		interaction	6.11	sig.	--	--
Chemistry, January 1949	I	sex	32.7	very sig.	boys	girls
		class	26.3	very sig.	AA	B
		interaction	3.86	sig.	--	--
	II	sex	8.67	very sig.	boys	girls
		class	7.34	very sig.	AA	B
		interaction	2.44	sig.	--	--
	I & II	sex	23.0	very sig.	boys	girls
		class	14.8	very sig.	AA	B
		interaction	1.5	not sig.	--	--
Chemistry, June 1949	I	sex	9.33	very sig.	boys	girls
		class	10.8	very sig.	AA	C
		interaction	1.0	not sig.	--	--
	II	sex	4.6	sig.	girls	boys
		class	9.00	very sig.	AA	C
		interaction	4.59	sig.	--	--
	I & II	sex	.02	not sig.	--	--
		class	11.25	very sig.	AA	C
		interaction	1.99	not sig.	--	--

TABLE VIII (Continued)

Examination	Part(s)	Source of Variance	F	Interpretation		
				Significance	High	Low
Chemistry, January 1950	I	sex	14.6	very sig.	boys	girls
		class	35.4	very sig.	AA	D
		interaction	2.20	not sig.	--	--
	II	sex	2.69	not sig.	--	--
		class	10.6	very sig.	AA	D
		interaction	3.35	sig.	--	--
	I & II	sex	6.85	sig.	boys	girls
		class	24.8	very sig.	AA	D
		interaction	1.12	not sig.	--	--
Chemistry, June 1950	I	sex	35.5	very sig.	boys	girls
		class	3.25	sig.	C	D
		interaction	5.28	sig.	--	--
	II	sex	3.89	not sig.	--	--
		class	4.06	sig.	B	C
		interaction	2.88	sig.	--	--
	I & II	sex	20.5	very sig.	boys	girls
		class	3.29	sig.	AA	A
		interaction	5.11	sig.	--	--
Earth Science, January 1949	I	sex	1.2	not sig.	--	--
		class	4.1	very sig.	A	C
		interaction	.8	not sig.	--	--
	II	sex	.26	not sig.	--	--
		class	.72	not sig.	--	--
		interaction	.96	not sig.	--	--
	I & II	sex	.42	not sig.	--	--
		class	.96	not sig.	--	--
		interaction	.45	not sig.	--	--

TABLE VIII (Continued)

Examination	Part(s)	Source of Variance	F	Interpretation		
				Significance	High	Low
Earth Science, June 1949	I	sex	17.0	very sig.	boys	girls
		class	5.43	very sig.	B	C
		interaction	2.26	not sig.	--	--
	II	sex	17.2	very sig.	boys	girls
		class	.8	not sig.	--	--
		interaction	1.25	not sig.	--	--
	I & II	sex	23.1	very sig.	boys	girls
		class	3.6	sig.	B	C
		interaction	.68	not sig.	--	--
Earth Science, January 1950	I	sex	1.56	not sig.	--	--
		class	2.63	not sig.	--	--
		interaction	1.14	not sig.	--	--
	II	sex	8.85	very sig.	girls	boys
		class	.94	not sig.	--	--
		interaction	1.34	not sig.	--	--
	I & II	sex	.65	not sig.	--	--
		class	1.60	not sig.	--	--
		interaction	.70	not sig.	--	--
Earth Science, June 1950	I	sex	1.42	not sig.	--	--
		class	1.25	not sig.	--	--
		interaction	.06	not sig.	--	--
	II	sex	.53	not sig.	--	--
		class	6.16	very sig.	A	B
		interaction	1.04	not sig.	--	--
	I & II	sex	.12	not sig.	--	--
		class	.80	not sig.	--	--
		interaction	.41	not sig.	--	--

TABLE VIII (Continued)

Examination	Part(s)	Source of Variance	F	Interpretation		
				Significance	High	Low
Physics, January 1949	I	sex	22.0	very sig.	boys	girls
		class	12.4	very sig.	AA	B
		interaction	1.94	not sig.	--	--
	II	sex	2.81	not sig.	--	--
		class	7.48	sig.	AA	A
		interaction	.25	not sig.	--	--
	I & II	sex	10.9	very sig.	boys	girls
		class	10.2	very sig.	AA	A
		interaction	.90	not sig.	--	--
Physics, June 1949	I	sex	.004	not sig.	--	--
		class	22.0	very sig.	AA	C
		interaction	1.92	not sig.	--	--
	II	sex	2.59	not sig.	--	--
		class	2.61	sig.	AA	C
		interaction	.6	not sig.	--	--
	I & II	sex	1.25	not sig.	--	--
		class	10.39	very sig.	AA	C
		interaction	1.56	not sig.	--	--
Physics, January 1950	I	sex	36.0	very sig.	boys	girls
		class	13.1	very sig.	AA	D
		interaction	2.36	not sig.	--	--
	II	sex	16.8	very sig.	boys	girls
		class	4.3	sig.	AA	B
		interaction	.9	not sig.	--	--
	I & II	sex	27.7	very sig.	boys	girls
		class	8.23	very sig.	AA	D
		interaction	.9	not sig.	--	--

TABLE VIII (Continued)

Examination	Part(s)	Source of Variance	F	Interpretation		
				Significance	High	Low
Physics, June 1950	I	sex	.02	not sig.	--	--
		class	5.03	sig.	A	C
		interaction	.40	not sig.	--	--
	II	sex	1.86	not sig.	--	--
		class	2.90	sig.	A	C
		interaction	.60	not sig.	--	--
	I & II	sex	.90	not sig.	--	--
		class	4.41	sig.	A	C
		interaction	.15	not sig.	--	--

total score of the Chemistry Examination for June 1950.

In these cases it is possible that differences in curriculum, and the organization and administration of the schools are the causes of observable variances, rather than the factors of size of school and sex.

However, in the cases where the observed variances could be attributed reasonably to sex or size of school, the following observations seem defensible:

1. On the various Biology Examinations, boys are significantly better than girls on two occasions, and girls are significantly better than boys on two occasions. Students from AA schools prove to be superior on nine occasions, while students from class D schools prove to be the lowest on five occasions, from class A on three occasions, and from class B on one occasion.

2. On the various Chemistry Examinations, boys are significantly better than girls on four occasions, while in no case did the girls prove to be significantly better than the boys. Students from the AA schools prove to be superior in five cases, while students from class B schools are lowest in one case, from class C in two cases, and from class D in two cases.

3. On the various Earth Science Examinations, boys are superior in three cases, while girls are superior in one. Students from the AA and A schools each prove to be superior in one case, and students from class B schools in two cases. Students from class C schools are lowest in two cases, and those from class B schools, once.

4. On the Physics Examinations, boys are superior in five cases, while in no case are girls superior. Students from the class AA schools are superior in eight cases, and those from class A schools, in three cases. Students from class A schools are lowest once, from class B twice, class C six times, and class D twice.

5. It may be stated, therefore, that in thirty-one out of forty-eight cases there is no variance attributable to the sex factor. However, out of the remaining seventeen cases the variances are significantly in favor of the boys in fourteen, and significantly in favor of the girls in three.

6. It may be stated that in eighteen cases there are no variances attributable to size of school. In the remaining thirty cases the variances are significantly in favor of students from class AA schools in twenty-three, while in no case are they in favor of students from class C or D. In nineteen of thirty cases students from class C and D schools appeared to exhibit less achievement, at least insofar as the variances may be criteria.

In conclusion, boys from the large high schools appear to score significantly higher on the Regents Examinations in Science than any other single group, while girls from small high schools appear to exhibit less achievement than any other single group.

SECTION V

THE VOCABULARY LOAD AND LEVEL OF READING DIFFICULTY OF THE REGENTS EXAMINATIONS IN SCIENCE[7]

The Problem

THE PROBLEM of this phase of the investigation is to evaluate the Regents Examinations in Science with respect to their vocabulary loads and levels of reading difficulty.

Methods Employed

The first step was to find a technique for evaluating the vocabulary load of the Regents Examinations in Science. The Flesch[8] formula (as well as other reading formulae) was obviously not suitable for the intended purpose since it is used ordinarily with passages of at least one hundred words or more and involves complete sentences rather than the type of material found on the Regents Examinations. Some of the examinations, for example, contain completion and multiple-choice items that do not adapt themselves readily to the use of the Flesch formula. Therefore it was decided to use the word-count method.

The first step was to tally all the words that appeared on the sixteen examinations. The words in the directions for writing the examinations however were not tallied, nor were numbers unless they appeared as words. Empirical formulas and structural formulas were not tallied. All other words, including those found on charts and diagrams were tallied.

Next, the words thus tallied were classified into two broad categories: (1) technical words, and (2) non-technical words. The technical category was further sub-divided into two classifications: (a) essential, and (b) desirable. The non-technical words were also divided into two classifications: (a) difficult, and (b) easy. These categories and classifications were established as follows: Letters were sent to sixty teachers who taught in each of the areas of Biology, Chemistry, Earth Science and Physics in the State of New York, asking if they would be willing to evaluate lists of vocabulary words in their respective teaching fields. A copy of this letter follows:

October 8, 1952

Dear

At the present time the University of the State of New York, through its science, research and statistical divisions, is undertaking an extensive investigation of the New York State Regents Examinations in Science. One of the facets of the investigation is to determine whether or not the vocabulary load on the examinations may be excessive.

According to the University, you have at one time or another taught Earth Science* and have administered and scored Regents Examinations in that area. Hence, we have a request to make of you. Would you be willing to evaluate a list of terms found on four representative Regents Examinations in Earth Science? If so, I would appreciate receiving an indication of your willingness on a postcard. If you agree, we shall do the following:

1. Send you a copy of the list of terms together with instructions and a self-addressed envelope in which to return your evaluation.
2. Give you full credit for your work in the final report as well as informing your administrator and board of education of your efforts.

The job will take about one hour and should be completed one week after receiving the material. We sincerely hope that you will have time to help us out.

Sincerely,

George G. Mallinson, Director
Evaluation Program for
Science Regents

lm

(*Note: The area of science named in the letter depended on the word list the individual was requested to evaluate.)

To this request, many teachers indicated their assent. They were then sent a mimeographed copy of all the terms that appeared on the examinations in their respective teaching fields. They were asked to evaluate each word on the list according to instructions found in an accompanying letter, a copy of which follows:

Dear

Your recent communication indicated that you would be willing to assist in evaluating the vocabulary content of the New York State Regents Examinations in Physics.* Your cooperation is more than appreciated.

Enclosed you will find a list of "Terms for Physics" together with a stamped envelope in which to return your evaluation. It is not necessary to sign your name. The following are the information and instructions for making the evaluation:

1. The list of terms consists of all the words and terms that appeared on the four successive Regents Examinations in Physics for January 1949, June 1949, January 1950 and June 1950.
2. The terms may be divided into two categories (a) non-technical, and (b) technical.
   (a) Non-technical terms are those that a person is likely to use at one time or another in his everyday conversation, or read in the newspaper or other literature not concerned specifically with physics.
   (b) Technical terms are those that a student would encounter specifically in a course in Physics. While such terms might be encountered in other courses or other places, an adequate understanding of the usual topics and principles of a typical course in physics would demand an understanding of, and the ability to use and apply, them.
3. Please examine the list of terms one by one. If you think a term fits the definition of "technical term", place an asterisk (*) before it. If you do not think the term fits the definition, ignore it. Reexamine your list to see if your judgment is consistent.
4. Then examine all the terms before which you placed an asterisk (*). If you believe that such a term is absolutely essential to an adequate understanding of topics and principles found in a typical course in physics, place a second asterisk (**) before the term. If however you believe the term to be merely a desirable technical term, leave it marked with but one asterisk (*).

Again let me say that your cooperation is absolutely essential and more than appreciated. In the final report due credit will be given your efforts and your administrator and board of education will be notified.

Your evaluation will be appreciated as soon as convenient and a copy of the final report will be sent you if you so request.

Sincerely,

George G. Mallinson, Director
Program of Evaluation
New York State Regents Examinations in Science

lm
Enc. 2

(*Note: The area of science named in the letter depended on the word list the individual agreed to evaluate.)

After a period of about four weeks, 23 Biology, 32 Chemistry, 24 Earth Science, and 24 Physics lists were returned. These were then tallied on a "master tally list" for each of the subjects.

If a word was checked "essential" by a total of ten or more respondents it was considered to be an "essential" term. (For example, the word "atomic" appearing on the list of Physics terms was checked "desirable" by five Physics teachers and "essential" by eleven.) However, if a word was checked "essential" or "desirable" by a total of five or more teachers (but checked "essential" by less than ten) it was considered "desirable." (For example the word "atmosphere" appearing on the list of Physics terms was checked "desirable" by seven Physics teachers, and "essential" by five. Therefore it was considered "desirable.")

All words rated as being "essential" or "desirable" were considered to be part of the technical vocabulary and hence were not deemed to be difficult. The remaining words, not rated as being part of the technical vocabulary, were considered to be non-technical terms, and therefore words which a student might find difficult. These non-technical words were then checked by means of the Buckingham-Dolch[9] word list in order to determine their grade-levels of difficulty. It was assumed that the courses in science would be taken by some students at these grade levels: Biology, ninth grade; Earth Science, tenth grade; Chemistry, eleventh grade; and Physics, eleventh grade. Any non-technical word was considered to be difficult therefore if it was rated above these respective grade levels in the word list.

Non-technical words not appearing in the Buckingham-Dolch list were also considered to be difficult.

Table IX lists the numbers of words on the various Regents Examinations in Science that fall into the various categories mentioned.

Conclusions

No listing will be made here of the different words falling into the various categories. However, insofar as the techniques used in this study may be valid, the following conclusions seem justified:

1. The greatest number of technical words (271) was found on the June 1950 examination in Biology, the fewest number (213) on the June 1949 examination in Physics. Thus the numbers of technical words on the different examinations does not vary greatly. Further it does not seem likely that the vocabulary load with respect to technical words is likely to be excessive.

2. The greatest number of difficult non-technical words (8) was found on the June 1950 examination in Biology, while there were no difficult non-technical words on the Chemistry examinations for January and June 1949. Hence it is rather unlikely that the numbers of difficult non-technical words are excessive.

3. The findings just indicated fail to show that there is any justification for criticizing the Regents Examinations in Science on the basis of their vocabulary loads and hence their levels of reading difficulty.

SECTION VI

ERRORS AND INCONSISTENCIES IN SCORING THE REGENTS EXAMINATIONS IN SCIENCE

The Problem

THE QUESTIONNAIRE that was sent to the science teachers in the initial stages of this investigation revealed that about two-thirds of them believed that a sampling of the examination papers should be rechecked after they were turned in to the State. This would seem to indicate that the teachers believed that there might be some errors and inconsistencies in the scoring of the examinations. Hence, it is the problem of this phase of the investigation (1) to determine whether or not the belief is valid; and if so, (2) to determine the types and frequencies of scoring errors that appear in the papers analyzed in this study.

Methods Employed

While tallying the scores that the students received on the various items on the examinations, the investigators recorded the obvious errors and inconsistencies that appeared in scoring them. The resulting lists were then studied. In general, it was found that the errors and inconsistencies could be classified into these four major categories:

1. Errors in the addition of points of credit
2. Errors resulting from failure to follow State-prescribed scoring procedures
3. Inconsistencies and errors in correction
4. Miscellaneous

Findings

1. Errors in the Addition of Points. —This category of error and inconsistency was by far the most extensive. There were several stages in the scoring of a paper where such errors in addition could occur, namely,

a) Errors in adding the Part I and Part II

TABLE IX

NUMBERS OF WORDS BELONGING IN DIFFERENT CLASSIFICATIONS

Field of Science	Date of Test	Number of Different Technical Words			Number of Different Non-Technical Words			Total Number of Different Words
		Essential	Desirable	Total Technical	Difficult	Easy	Total Non-Technical	
Biology	Jan. 1949	143	121	264	1	336	337	601
	June 1949	134	129	263	4	415	419	662
	Jan. 1950	119	135	254	4	363	367	621
	June 1950	139	132	271	8	385	393	664
Chemistry	Jan. 1949	171	73	244	0	184	184	428
	June 1949	183	59	242	0	190	190	432
	Jan. 1950	188	67	255	1	190	191	446
	June 1950	174	72	246	1	199	200	446
Earth Science	Jan. 1949	110	126	236	3	217	220	456
	June 1949	110	145	255	3	238	241	496
	Jan. 1950	110	130	240	6	217	223	463
	June 1950	97	145	242	5	228	233	475
Physics	Jan. 1949	108	130	238	6	319	325	563
	June 1949	105	108	213	6	323	329	542
	Jan. 1950	101	133	234	4	333	337	571
	June 1950	111	129	240	4	327	331	571

scores to obtain the total test score.

b) Errors in totaling the Part I score. There were two major chances for error here; a mistake could be made (1) in totaling the number of points to be deducted because of a student's failure to answer items correctly; and (2) in subtracting this number of points from the maximum point value of fifty for Part I.

c) Errors in totaling the Part II score. In this case there were several ways in which the errors could occur. At times simple errors could be made in totaling the scores on the parts of items. At other times errors in subtraction resulted when the points to be deducted because of error were totaled and subtracted from ten (the maximum point value for each individual Part II item). In still other cases the correction marks of the teacher were so light or illegible that they were apparently overlooked when the points were totaled. (This latter type of error would, of course, have been avoided if cumulative scores had been kept, as suggested by the State.)

Certain other errors were made in totaling Part II scores because of the failure to follow correct scoring procedures. Such cases will be discussed in the next two parts of this section.

It is interesting to note that the greatest number of errors in making totals accrued to the benefit of the student, that is, the total score awarded the paper was higher than the correct total. For example, out of 2011 Biology Examinations for January 1949, ninety-two errors in making totals were detected. Of these, only six scores were lower than they should have been; the rest were higher.

The tables that follow indicate the frequency with which the various types of errors just described occurred. Since the approximate percentages of errors were the same for all sixteen sets of examinations, data are included for only one examination in each of the four areas of science.

2. Errors Resulting from Failure to Follow State-Prescribed Scoring Procedures. —The State of New York issues a manual[10] listing the procedures to be followed in scoring Regents Examinations. Many of these suggested procedures were violated by a number of teachers, and these violations in many instances led to incorrect examination scores. Examples of these types of errors follow:

a) Failure to keep a cumulative score. The State suggests that for each item or part of an item the points awarded should be indicated on the test paper and a cumulative positive score be kept throughout the paper. This means that the points awarded for answering correctly each item or part of an item be totaled continuously as the paper is scored. Hence at the completion of the last item the resulting cumulative score will represent the total score of the paper.

By far the greatest number of teachers tallied the points deducted rather than the points awarded for the answers to items or parts of items. Among those who did indicate the points awarded, many failed to keep cumulative scores. Obviously the practice of indicating the number of points deducted is more subject to error than the method recommended by the State. The teacher may make a mistake in totaling the points to be deducted and make another in subtracting this total from the maximum point value allotted the item or part of an item. For example, such errors occurred on 117 out of 1699 papers in Earth Science for June 1949.

b) Failure to score items in sequence. On Part II of the examinations a student has the option of selecting five out of a possible eight or nine items. On some of these items he may omit one or two parts of the item. Occasionally all eight or nine items or all parts of a single item were answered by a student. In certain cases if one item (or part of an item) that appeared in the middle was answered poorly or incorrectly, some teachers skipped it and gave credit for the later sections that were answered more correctly.

The State requires that the items or parts of items should be scored in order of appearance, omitting the last item or part. Failure to do so, of course, may give a student a higher score than he deserves.

c) Scoring of papers by several different teachers. The State suggests that one teacher should score all items on any given examination paper. However, in many cases, particularly in larger schools, it was obvious that teachers worked together on scoring a group of papers. For example, one teacher might score items one through ten on a group of papers; another teacher items eleven through twenty. Such a procedure often lead to inconsistencies and errors when the total score of a paper computed.

d) Counting scores below 62 as passing. The cutting score for passing papers has been set by the State at 65. However, recognizing that errors may occur in the correction of papers, a three percent correction error has been allowed. Thus scores from 62 to 64 are considered as "passing". These "below level" scores of 62, 63 and 64 are recorded as (65). In several cases, however, scores of 61, 60 and even 59 and 58 were recorded as (65).

In addition, scores that should have been recorded as (65) were sometimes recorded as 65. This occurred on twenty-four papers on the Bi-

TABLE X

NUMBERS AND PERCENTAGES OF ERRORS IN THE ADDITION OF POINTS ON THE REGENTS EXAMINATIONS IN BIOLOGY FOR JUNE, 1949

Score	Total No. of Papers	Number of Errors	Percentage of Errors	Score	Total No. of Papers	Number of Errors	Percentage of Errors
(65)	280	44	15.6%	82	62	6	9.7%
65	62	13	21.0%	83	64	10	15.6%
66	58	11	19.0%	84	64	8	12.3%
67	61	8	13.0%	85	62	10	16.0%
68	48	6	12.5%	86	53	7	13.0%
69	64	7	11.0%	87	30	6	20.0%
70	89	9	10.0%	88	51	4	7.8%
71	48	8	16.7%	89	42	6	14.3%
72	66	13	19.6%	90	42	4	9.5%
73	60	13	22.0%	91	38	2	5.3%
74	54	8	14.8%	92	38	8	21.0%
75	100	32	32.0%	93	32	5	15.6%
76	57	8	14.0%	94	21	2	9.5%
77	84	12	14.3%	95	12	2	16.7%
78	87	19	21.8%	96	12	2	16.7%
79	67	6	9.0%	97	13	2	15.4%
80	82	10	12.2%	98	7	0	0
81	83	8	10.6%	99	1	0	0
				100	0	0	0

TABLE XI

NUMBERS AND PERCENTAGES OF ERRORS IN THE ADDITION OF POINTS ON THE REGENTS EXAMINATIONS IN CHEMISTRY FOR JANUARY, 1950

Score	Total No. of Papers	Number of Errors	Percentage of Errors	Score	Total No. of Papers	Number of Errors	Percentage of Errors
(65)	105	7	6.7%	82	61	6	9.8%
65	37	7	18.9%	83	64	6	9.4%
66	33	6	18.2%	84	64	6	9.4%
67	35	6	17.2%	85	64	0	0
68	36	6	16.7%	86	69	3	4.4%
69	36	6	16.7%	87	69	3	4.4%
70	33	2	6.1%	88	75	8	10.7%
71	47	1	2.1%	89	61	4	6.6%
72	37	3	8.1%	90	73	5	6.8%
73	40	8	20.0%	91	67	3	4.5%
74	41	8	19.5%	92	79	7	8.9%
75	64	7	10.9%	93	73	3	4.1%
76	51	8	15.7%	94	75	0	0
77	59	3	5.1%	95	72	2	2.8%
78	46	1	2.2%	96	72	1	1.4%
79	47	3	6.4%	97	75	1	1.3%
80	66	5	7.1%	98	57	1	1.8%
81	55	5	9.1%	99	60	1	1.7%
				100	47	2	4.2%

TABLE XII

NUMBERS AND PERCENTAGES OF ERRORS IN THE ADDITION OF POINTS ON THE REGENTS EXAMINATIONS IN EARTHSCIENCE FOR JANUARY, 1950

Score	Total No. of Papers	number of Errors	Percentage of Errors	Score	Total No. of Papers	Number of Errors	Percentage of Errors
(65)	163	26	15.9%	82	44	6	13.6%
65	50	12	24.0%	83	40	4	10.0%
66	40	6	15.0%	84	41	6	14.6%
67	46	7	15.2%	85	43	6	13.9%
68	51	8	15.7%	86	38	3	7.9%
69	62	8	12.9%	87	29	1	3.4%
70	47	5	10.7%	88	30	4	13.3%
71	50	9	18.0%	89	30	2	6.7%
72	8	3	37.5%	90	31	2	6.4%
73	1	0	0	91	22	1	4.1%
74	0	0	0	92	14	1	7.1%
(75)	158	16	10.1%	93	15	1	6.7%
75	62	15	24.2%	94	9	2	22.2%
76	38	5	13.2%	95	9	0	0
77	47	4	8.5%	96	7	1	14.3%
78	43	6	13.9%	97	2	1	50.0%
79	48	6	12.5%	98	1	0	0
80	47	3	6.4%	99	0	0	0
81	53	6	11.3%	100	0	0	0

TABLE XIII

NUMBERS AND PERCENTAGES OF ERRORS IN THE ADDITION OF POINTS ON THE REGENTS EXAMINATIONS IN PHYSICS FOR JANUARY, 1950

Score	Total No. of Papers	Number of Errors	Percentage of Errors	Score	Total No. of Papers	Number of Errors	Percentage of Errors
(65)	206	14	6.8%	82	65	6	9.2%
65	71	11	15.5%	83	64	3	4.7%
66	71	9	12.7%	84	74	3	4.1%
67	64	4	6.2%	85	82	9	10.9%
68	69	8	11.6%	86	65	1	1.5%
69	48	8	16.7%	87	59	9	15.3%
70	74	7	9.5%	88	58	4	6.9%
71	60	6	10.0%	89	64	3	4.7%
72	63	14	22.2%	90	60	2	3.3%
73	70	6	8.6%	91	62	4	6.4%
74	66	9	13.6%	92	46	1	2.2%
75	88	9	10.4%	93	48	1	2.1%
76	71	9	12.7%	94	36	0	0
77	56	2	3.6%	95	24	1	4.2%
78	75	7	9.3%	96	36	1	2.8%
79	55	2	3.6%	97	30	0	0
80	103	8	7.8%	98	16	0	0
81	77	8	10.4%	99	5	0	0
				100	2	0	0

ology Examination for January 1950.

e) Counting scores of 72, 73, and 74 as (75). In some cases, such as in schools that do not offer laboratory work, and in "short-term" courses (veterans' classes) a score of 75 is required for a passing grade. In many cases teachers counted 72, 73 or 74 as (75), similar to counting 62-64 as (65). While this is not a legal procedure, the State has accepted it in the past, considering such scores as falling within a scoring-error range similar to the 62-65 range. However, in a number of cases in which 75 was not required for passing, the teachers still recorded 72, 73 or 74 as (75).

3. Errors and Inconsistencies in Correction. —The summaries made by the investigators of scoring errors revealed several types of errors that could be classified as being actual mistakes or inconsistencies in scoring the papers. Examples of such errors follow:

a) Giving credit for parts of an item that should have been rejected. As indicated previously, many Part II items contained seven or eight parts, of which the student was to answer five or six and omit the remainder. Sometimes a scorer would fail to note that all parts of such an item had been answered by a student, would give credit for all parts, and hence award more than the maximum allowable number of points to the student. Such an error occurred in twenty-two out of 2441 Biology Examinations for January 1950.

b) Giving credit for an entire item that should have been omitted. An error similar to (a) above occurred when a student had the option of rejecting one or more entire items in Part II. Occasionally a student would answer all the items, and again the scorer would correct all of them, giving the student extra points.

c) Failure to note the omission of entire items or parts of items. A condition just the reverse of cases (a) and (b) above occurred when a teacher did not notice that a student had omitted an entire item or part of an item. This type of error occurred quite easily if the teacher was scoring the paper by deducting points (rather than recording the number of positive points awarded, as suggested by the State). The teacher would simply add the number of points deducted and subtract from the maximum number of points allotted the item, failing to note the omission. Had the paper been scored by the positive-point method, such an error would not have resulted. Errors of this type were found on twenty out of 1699 papers in Earth Science for June 1949.

d) Failure to correct an item or part of an item. Occasionally a teacher would overlook an entire item or part of an item. In such a case the student did not receive as high a score as he deserved. This type of error was detected on fifty-nine out of 2441 Biology Examinations for January 1950.

e) Errors in awarding points. The items on Part II of the examinations consist of from two to eight or nine parts. The maximum point value of the items is in each case ten. However, the values of the parts vary from item to item, depending on the number of parts in the item or the complexity of the part. For example, one item might consist of three parts of values five, three, and two respectively; while another item might consist of five parts of two values each. In many cases the teachers awarded incorrect numbers of points for items, namely, parts were given three points credit, when the maximum value was only two.

The opposite situation, the awarding of too few points, was difficult to detect. However, one bundle of sixty Physics papers was tallied on which the teacher did not give full credit (three points) in any case. It is difficult to believe that all sixty students failed to answer the item correctly. Hence it seems reasonable to assume that the teacher thought that the maximum value of the item was two, rather than three.

f) Inconsistencies in scoring. Many inconsistencies in scoring were noted, not only within the work of a single teacher, but also between the scoring procedures of different teachers. An answer that would receive full credit from one teacher might receive only partial credit or no credit from another. Such situations were expecially obvious on items that required drawings or diagrams. Often one teacher would apparently give more credit for neat, artistic work, while another would consider only the scientific accuracy of the drawing.

Similar inconsistencies occurred even within the work of one teacher. For example, one Physics Examination included an item involving unites of electricity. One teacher gave credit for an answer of 2520 watts, but consistently marked as incorrect answers of 2.52 kilowatts. Since the students were given no instructions in the item as to the units to be used, it would appear that both answers should be considered equally correct.

g) Obvious errors in scoring. In many cases errors were detected (1) where a correct answer was consistently scored by a teacher as being wrong, and (2) where obviously incorrect answers were marked correct. It may be assumed that in such cases the teacher simply did not know the correct answer to the item.

4. Miscellaneous Errors in Scoring. —Certain types of errors appeared that were difficult to classify under any of the previous categories. Hence they were grouped under this heading. They are as follows:

a) Errors in transposition. In most cases

the examination papers of the individual students are stapled to a cover page on which the teacher lists the total Part I score, the scores awarded on the individual Part II items, and the Part II total score. These are then totaled on this cover sheet to show the student's final score on the entire examination.

Many times errors were made in the transfer of these partial scores to the cover page. For example, such errors occurred on 118 out of 1974 papers for the Chemistry Examination for June 1949.

It is difficult to determine whether these errors are due to the carelessness of the scorer, or whether they are intentional. However, it is interesting to note that most of these errors accrued to the benefit of the student, that is, the score on the cover was higher than the actual score. Occasionally, however, a scorer would fail to record the score obtained on an entire item on Part II. Thus the student received a lower score than he earned.

b) Obvious "upgrading" of scores. In some cases it was quite evident that the scorer had remarked a paper to give a student a higher grade than he earned or had simply changed the total grade to bring it up to a passing score. Such cases are difficult to construe as anything but dishonesty on the part of the scorer. For example, twenty-two such papers were identified in the Chemistry Examination for June 1949.

c) Failure to grade a paper completely. A situation somewhat related to (b) was an obviously dishonest practice that was detected on a few occasions. In several instances papers were found in which the scorer had corrected Part I of the paper only. The scorer then simply credited the paper with a sufficient number of points on Part II to bring the total up to the passing level. In some cases the Part I score was merely doubled, or if this did not give the paper a passing score, a sufficient number of additional points was added. It was obvious that the items on Part II had not even been read in some of these cases.

Recommendations

As a result of these findings the following recommendations seem reasonable:

1. It is recommended that the State of New York provide a more specific list of instructions for the scoring procedures to be followed in correcting the Regents Examinations in Science.
2. It is recommended that a scoring key be provided for Part II of the examinations, as well as for Part I. It is realized that because of the nature of the Part II items, it is difficult to construct an absolute scoring key. However, it would seem desirable to prepare a "scoring guide" that would suggest point values for whole or partial answers.
3. It is recommended strongly that the State continue to spot-check the examination papers in an attempt to identify scoring errors and inconsistencies, and that, further, they notify the science supervisor or administrators in the schools in which the errors most frequently occur of their type and extent. This may reduce the appearance of these errors on future examinations.

SECTION VII

AN ANALYSIS OF THE SCORES OBTAINED ON THE TEST ITEMS ON THE REGENTS EXAMINATIONS IN SCIENCE

The Problem

THE PROBLEM of this phase of the investigation is to analyze the individual items on the sixteen Regents Examinations in Science that were studied, with respect to these points:

1. The degree of difficulty of the various types of items
2. The discriminating power of the items
3. The popularity of certain items

Methods Employed

In order to determine the degree of difficulty and the discriminating power of the individual items of the examinations, the average or percentage score obtained on each item was tabulated as described below. For Parts I of the tests (each of which are composed of fifty short-answer type items of unitary value) the average score for each item was determined for each of the total score groups (65 through 100) by dividing the number of students answering the item correctly by the total number of students receiving the respective score.

As has been stated, Parts II of all the examinations are composed of eight or nine essay-type items each of which bears a total value of ten points. Of these, the student may select any five. Each essay item consists of from two to ten parts of varying point values. To calculate the percentage scores on these parts, the total number of points earned by all the students answering the part was divided by the maximum number of points that they could have obtained had they all answered it correctly. This was done for each item for each total score group from 65 through 100.

Degrees of Difficulty of the Items

The average and percentage scores thus obtained were analyzed to determine the degree of difficulty of each item on each of the sixteen examinations. It is obvious that if the average or percentage score of any item was consistently low for all the total score groups, the item must be difficult. Conversely, if the score was consistently high, the item must be easy. The items were categorized arbitrarily as being "easy," "of average difficulty," or "difficult" on the basis of the following criteria: (1) if approximately one-half or more of the average or percentage scores were above .90, the item was considered easy; (2) if approximately one-third or more were below .50, the item was considered difficult; and (3) those falling between these ranges were considered to be of average difficulty.

The items thus categorized were then studied in an effort to determine whether any one type seemed to be consistently easy or difficult. As a result of this analysis, a list was made that included the general subject-matter areas containing the largest numbers of difficult items on the examinations in each of the science fields. A few examples of each type are cited below:

Difficult Items

I. Biology
  A. Plant and Animal Phyiology
    1. "What are two adaptations of a root hair that help it to perform its functions?" (January 1950, Part II, 1 a_2)

    2. "Name a plant tissue cell. State its special function and describe how it is fitted to perform this function." (June 1950, Part II, 7 b)

    3. "The mouth waters when food is present because salivary glands have been stimulated by _____ neurons." (January 1949, Part I, 34)

  B. Genetics and Heredity
    1. "To the species involved, mutations are (1) always harmful, (2) always useful, (3) usually harmful, (4) usually useful." (January 1950, Part I, 13)

    2. "Give an explanation for the following true statement(s): In some cases the marriage of first cousins results in very desirable offspring; in other cases very undesirable offspring result." (January 1950, Part II, 3 f)

    3. "An animal has four chromosomes in each body cell. State the number of chromosomes in (1) a primary egg cell." (January 1950, Part II, 5 b_1)

  C. General Terminology
    1. "The process of boiling milk to kill all bacteria is sterilization." (January, 1950, Part I, 50)

    2. "The part of the seed that will develop into the plant is the _____." (January 1949, Part I, 38)

    3. "An example of an antibiotic is sulfadiazine." (June 1950, Part I, 25)

There were also moderate numbers of difficult items in the general categories of comparative anatomy, bio-chemistry, and history of biology.

II. Chemistry
  A. Organic Chemistry
    1. "Hard coal consists chiefly of (1) carbohydrates, (2) combined carbon (3) uncombined carbon, (4) hydrocarbons." (January 1950, Part I, 17)

    2. "Describe one method of making methyl alcohol." (January 1950, Part II, 6 e_1)

    3. "Write the structural (graphic) formula for (1) chloroform, (2) ethylene." (June 1949, Part II, 6 b)

  B. Atomic Weights
    1. "The weight of 22.4 liters of hydrogen is approximately (1) 0.09 grams (2) 2 grams, (3) 1 gram, (4) 22.4 grams." (January 1949, Part I, 33)

    2. "The weight of nitrogen compared with an equal volume of air is approximately (1) one-half as great, (2) the same, (3) twice as great, (4) fourteen times as great." (June 1950, Part I, 47)

  C. Commercial Reactions
    1. "Charcoal is a product of the process that also produces (1) acetic acid (2) coal tar, (3) coke, (4) gasoline." (January 1950, Part I, 16)

    2. "The reaction of carbon monoxide and hydrogen is used commercially to make (1) carbonic acid, (2) chlorine, (3) methanol, (4) soap." (June 1950, Part I, 27)

    3. "Name two products which are obtained from coal tar." "State one use for each product mentioned in c." (June 1949, Part II, 6 c, d)

4. "What substance may be treated with chlorine to manufacture bleaching powder?" (June 1950, Part II, 3 e)

D. Laboratory Procedures and Techniques
1. "If too much air is allowed in the fuel mixture, the Bunsen flame will (1) become colorless, (2) become yellow, (3) deposit soot, (4) strike back." (June 1949, Part I, 28)

2. "Give the reagents used in the laboratory preparation of (1) nitric acid (2) ammonia." (June 1949, Part II, 8 d)

3. "State briefly how to prepare hydrogen from water and sodium chloride." (June 1950, Part II, 5 d_2)

E. Equations
1. "Write a completely balanced equation for the reaction between copper and hot concentrated sulfuric acid." (June 1950, Part II, 1 e)

2. "Write an ionic equation to show what happens when an oxygen ion is converted to an O_2 atom." (June 1949, Part II, 4c)

Other difficult items include those involving terminology, characteristics of elements and compounds, and everyday applications of chemistry.

III. Earth Science
A. Geology (this category included by far the largest number of difficult items)
1. "Headwater erosion of a valley glacier results in the formation of a (an) _____." (January 1950, Part I, 2)

2. "Explain the following true statement(s): The Catskill Mountains are classified as a plateau region." (January 1950, Part II, 5 d)

3. "Explain how weathered rock may again become bedrock." (January 1949, Part I, 1 d)

4. "An intrusion of igneous rock that cuts across the rock layers is called a (1) dike, (2) fault, (3) laccolith, (4) sill." (June 1950, Part I, 31)

B. Weather
1. "Distinguish between absolute and relative humidity." "Explain why relative humidity decreases as temperature increases." (January 1949, Part II, 2 c, d)

2. "Air descending the side of a mountain becomes compressed. Why does this make the air comparatively dry?" (January 1950, Part II, 6 c)

3. "Barometric pressure recorded on a weather-bureau station model as 247 would be read (1) 924.7, (2) 1002.47, (3) 1024.7, (4) 1247 millibars." (June 1950, Part I, 23)

C. Astronomy
1. "The planet which is about the same size as the earth is (1) Mars, (2) Venus, (3) Mercury, (4) Uranus." (June 1949, Part I, 29)

2. "Explain the following:
In New York State the altitude of the noon sun is higher during the summer than it is during the winter." (January 1949, Part II, 5 b)

IV. Physics
A. Sound
1. "State two conditions under which two sound waves of the same amplitude will produce complete interference." (January 1950, Part II, 6 c; and June 1950, Part II, 4 d)

2. "The note produced by a string vibrating as a whole is called a (an) overtone." (January 1949, Part I, 38)

3. "Find the fundamental frequency in vps. of a note produced by a whistle, closed at one end, if the length of the air column is six inches. Air temperature is 20 degrees, C." (June 1950, Part II, 4 c)

B. Electricity
1. "During the discharging process of a lead storage cell, the amount of water in the cell _____." (June 1949, Part I, 44)

2. "An electric heater has two coils with resistances of 40 ohms and 60 ohms. The heater operates on a 120-volt circuit. It is equipped with a switch that allows either coil to operate in series." "In which of the three possible operating circuits is the heat developed the greatest?" (January 1950, Part II, 4 d)

3. "An iron wire has more resistance

than a copper wire of the same dimensions, and an aluminum wire has more resistance than the copper wire of the same dimensions. Compare the current in the three wires and state in which wire the most heat is generated when they are connected to a battery (1) in series; (2) in parallel." (June 1950, Part II, 5 b)

C. Lenses and Mirrors
1. "The image of an object viewed through a concave lens is always erect and larger than the object." (June 1949, Part I, 26)

2. "A woman sees a full-length image of herself in an upright plane mirror. The minimum length of the mirror is (1) exactly the same as, (2) one-half, (3) twice, (4) independent of the height of the woman." (January 1950, Part I, 12)

Other physics items that seemed difficult occurred in the areas of heat and mechanics.

It is interesting to note that the total number of items categorized as being "easy" far outnumbered those categorized as "difficult." This, of course, is explained partly by the fact that the examinations analyzed all received passing scores between 65 and 100. Hence, the examination items would naturally consist chiefly of those receiving high average and percentage scores.

Easy Items

I. Biology
A. Conservation
1. "Contour plowing is done in an effort to (1) beautify the farm, (2) control weeds, (3) discourage insects, (4) save topsoil." (January 1949, Part I, 6)

2. "Explain the relationship of forests to each of the following: (1) flood control, (2) prevention of erosion, (3) preservation of wildlife." (January 1950, Part II, 1 b)

B. Plant and Animal Physiology
1. "In mammals the body wastes are excreted by the lungs, skin and (1) kidneys, (2) pancreas, (3) small intestine, (4) stomach." (January 1949, Part I, 21)

2. "The type of cell in the bloodstream that increases in number in response to the invasion of bacteria is the _____." (June 1950, Part I, 36)

3. "State four life functions carried on by a maple tree." (January 1949, Part II, 1 a)

C. Reproduction and Genetics
1. "An important function of the sperm cell is to supply the egg with (1) a set of genes, (2) extra cytoplasm (3) extra food, (4) important hormones." (January 1949, Part I, 7)

2. "The union of two unlike sex cells is called (1) fertilization, (2) maturation, (3) parthenogenesis, (4) vegetative propagation." (January 1950, Part I 3)

3. "Using a keyed and labelled diagram, show the cross between long and long radishes." (January 1950, Part II, 9 c_1)

D. Everyday Applications of Biology
1. "It is now possible to keep an area quite free from flies by the use of (1) 2-4D, (2) DDT, (3) Streptomycin, (4) sulfa drugs." (January 1949, Part I, 5)

2. "State the principal health purpose of each of two of the following: chest x-rays; Wassermann test; pasteurization of milk." (January 1949, Part II, 6 c)

II. Chemistry
A. Chemical and Physical Properties of Elements and Compounds
1. "Hydrogen sulfide is most easily recognized by its (1) color, (2) density, (3) odor, (4) state." (January 1949, Part I, 1)

2. "The lightest of the following gases is: (1) NH_3, (2) NO, (3) N_2O, (4) NO_2." (June 1949, Part I, 1)

B. Chemical Reactions
1. "The solution resulting from the reaction between sodium and water contains (1) an acid, (2) an anhydride, (3) a base, (4) a salt." (January 1949, Part I, 10)

2. "The reaction of a carbonate with an acid yields (1) carbon dioxide, (2) carbon monoxide, (3) hydrogen, (4) oxygen." (June 1949, Part I, 16)

3. "Give three reasons why a chemical reaction may go to completion." (January 1950, Part II, 4 d)

C. Everyday Applications of Chemistry

1. "The growth of a legume, such as clover, adds to the soil a compound of (1) nitrogen, (2) phosphorous, (3) potassium, (4) sulfur." (January 1949, Part I, 22)

2. "Goiter may be caused by a diet deficient in (1) bromine, (2) chlorine, (3) flourine, (4) iodine." (January 1950, Part I, 25)

D. Laboratory Procedures

1. "To prepare bromine in the laboratory, add sulfuric acid to (1) NaBr, (2) NaBr and MnO_2, (3) NaCl and Na Br, (4) $MnBr_2$." (June 1949, Part I, 37)

2. "A catalyst used in a preparation of oxygen is (1) manganese dioxide, (2) mercuric oxide, (3) potassium chlorate, (4) potassium chloride." (January 1950, Part I, 33)

3. "Draw a diagram of the apparatus used in preparing and collecting ammonia in the laboratory." (January 1949, Part II, 5 c)

Other easy items in Chemistry included those involving the writing of balanced equations, and the knowledge of symbols and formulae.

III. Earth Science

A. Geology

1. "Physical and chemical action on exposed rock surfaces by atmospheric agencies is called (1) erosion, (2) corrosion, (3) suspension, (4) weathering." (June 1949, Part I, 23)

2. "The breaking of minerals in such a way that smooth plane surfaces are produced is known as (1) cleavage, (2) fracture, (3) luster, (4) streak." (January 1949, Part I, 17)

3. "The peeling or splitting-off of outer layers of rock due to temperature changes is called (1) cleavage, (2) exfoliation, (3) faulting, (4) fracture." (June 1950, Part I, 35)

B. Weather

1. "Closely spaced isobars on a weather map indicate _____ winds." (January 1949, Part I, 8)

2. "When the air is completely saturated with moisture, the relative humidity is zero %." (June 1949, Part I, 38)

3. "State two characteristics of weather that an mT air mass will bring to New York State." (June 1950, Part II, 2 b)

4. "Distinguish between weather and climate." (January 1950, Part II, 6 a)

IV. Physics

A. Electricity and Magnetism

1. "The filament now used in most electric lamps is made of _____." (January 1950, Part I, 24)

2. "When the south pole of a magnet is brought near the head of an iron nail, the head of the nail becomes a south pole." (June 1950, Part I, 33)

3. "A step-up transformer used to operate a neon sign has a turn ratio of 1:100. The primary voltage is 110 volts. The primary current is 10 amperes. The secondary current is .09 ampere. Find the (b) wattage of the primary; (c) wattage of the secondary." (January 1949, Part II, 6 b, c)

B. Mechanics

1. "The moment of a 20-pound force pushing perpendicularly on a lever five feet from the fulcrum is _____ pound-feet." (June 1949, Part I, 20)

2. "The theoretical mechanical advantage of a wheel and axle is 6. The wheel diameter is 12 inches. The axle diameter is _____ inches." (January 1950, Part I, 20)

3. "A 500 pound weight is drawn up an inclined plane 15 feet long and 3 feet high. The effort required is 125 pounds. Find the actual mechanical advantage." (June 1950, Part II, 3 a)

C. Density

1. "As a liquid contracts, its density _____." (June 1949, Part I, 37)

2. "Two solids show equal apparent losses of weight when submerged in water. Their densities must be equal." (June 1950, Part I, 38)

3. "The apparent weight of 3 cubic feet of metal submerged in water is 375 pounds. (Density of water is 62.5 pounds per cubic foot). Find the (a) volume of water displaced; (b) weight of water displaced;

(c) weight of metal in air." (January 1949, Part II, 1 a, b, c)

Other easy items in Physics included many in the areas of light, sound, and the use of scientific instruments.

Following the analysis of the difficulty of the items as a function of subject-matter, an analysis was made to determine the relationship between the forms in which the items were written and their degrees of difficulty.

Parts I of all the examinations are composed of short-answer type items such as the multiple choice, modified true-false, and completion types. Parts II of the examinations are more subjective in nature, and include essay-type items that require more explanations and descriptions; the drawing or interpretation of diagrams; mathematical problems; and the writing of equations.

On the Biology Examinations, it was found that the largest percentage of the difficult items on Part I were of the completion type, while the easiest were the multiple-choice. On the Part II items, no particular type of item seemed easier or more difficult than the others.

The only type of short-answer type of item found on the Chemistry Examinations was the multiple-choice type. Hence, no comparison could be made. On Part II, however, the highest percent of the difficult items included those that required explanations, and those that demanded the writing of equations.

On Part I of the Earth Science Examinations, the modified true-false and the completion items seemed to be about the easiest for the students; while the multiple-choice seemed to be the most difficult. On Part II the percentage of difficult essay items was high, but among the easy items were those that required drawings or the interpretation of diagrams.

On Parts I of the Physics Examinations, the three types of short-answer items were of approximately equal difficulty. On Parts II, however, the number of easy mathematical items was high. A number of mathematical items seemed to be difficult, also.

An overall comparison of the items on Parts I and II of all the examinations indicates that there is a substantially higher percentage of difficult items on Part II than on Part I. Hence, it appears that in general, the short-answer type items are less difficult for most students than are the "essay-type." However, the data just summarized fail to reveal any consistent trends concerning the degrees of difficulty of the various types of items. Hence, the specific form in which the item is written does not generally appear to be a significant factor in its degree of difficulty.

Summary

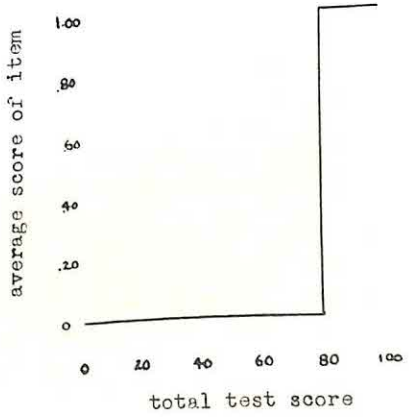
From an analysis of the degrees of difficulty of the various examination items, a few generalizations may be made:

1. It appears that items involving current science information (such as antibiotics, the hydrogen bomb, etc.,) seem to be more difficult than other types. A possible explanation may be the fact that many textbooks are not up-to-date.
2. Many of the difficult items are also ambiguous, and hence present difficulty for the student in answering, as well as problems of scoring for the teacher.
3. Items involving the use of scientific attitudes, applications of knowledge, and the use of elements of scientific method are in general more difficult than the factual type.
4. Essay-type items are generally more difficult than the short-answer items.

Discriminating Power of the Examination Items

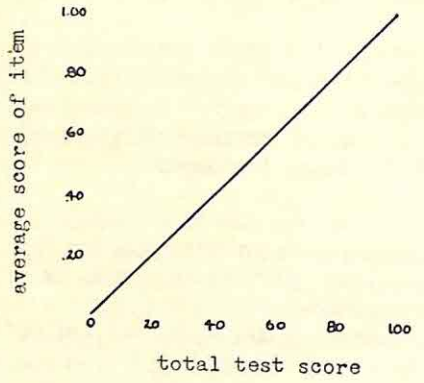
In general, there are two different views with respect to the concept of discriminating power of an examination item:

1. If an item is designed to measure one precise objective, it will ideally cut the examination group into two sections, namely, those students above a given total score who answer the item correctly, and those below the given total score who answer the item incorrectly. If the average scores for such an ideal item were plotted, the resulting histogram would have the following configuration:



2. If an item is designed to measure a generalized objective, or a multiple set of objectives; or if the item is used to test a group of individuals whose quality and quantity of training with respect to the objectives differ, then

an analysis of the average scores of the item would ideally rise gradually as the total scores of the group increase. A graph similar to the following would result:



The New York Regents Examinations in Science are designed to measure a multiple set of objectives. In addition, although guided by a State Syllabus, no teacher is required to present a given course of study. Hence, no two teachers are likely to teach the same kind or amount of science material. Therefore no two groups of students who take the examinations are likely to have the same training. For this reason, the discriminating power of the examination items needs to be evaluated in terms of the second viewpoint described above.

Thus, the average scores of the items on Parts I of the examinations and the percentage scores of the items on Parts II were analyzed to determine whether they increased as the total scores increased from (65) to 100. The following criteria were established for use in this study, as a measure of discriminating power:

1. An item was considered to have excellent discriminating power, if (after the initial increase) its average score, in seventy-five percent of the cases, increased consistently (a one to ten percent increase) with each one point increase in the total examination score.
2. An item was considered to have moderate or "average" discriminating power if the increase in twenty-five percent or more of its average or percentage scores, fluctuated between ten and twenty-five percent.
3. An item was considered to be a poor discriminator if the increase in twenty-five percent or more of its average or percentage scores fluctuated by more than twenty-five percent.
4. All items that had consistently low or consistently high average or percentage scores were also considered to have poor discriminating power.

Results

Table XIV summarizes the findings of the analysis for the discriminating power of the examination items.

Table XIV indicates that on the Biology Examinations, the greatest number (approximately sixty percent) of the items was found to have average discriminating power, while only about four percent were found to be excellent discriminators. Of the poor discriminators, four percent were so classified because of the great fluctuation of their average scores. Approximately thirty percent were considered poor because they were easy, and two percent because they were difficult. It is interesting to note that of the easy items the vast majority were found on Part I.

There were only three (about one percent) of the Chemistry items that could be classified as excellent discriminators on the basis of the criteria outlined above. Again, the majority of the Chemistry items (about forty-three percent) were found to be of average discriminating power. Of the poor discriminators, twenty-four percent were so classified because their average scores fluctuated greatly; about thirty percent were easy, and about two percent difficult. Again, the majority of the easy items appeared on Part I of the examinations.

On the Earth Science Examinations, about two percent of the items were considered as being excellent discriminators while approximately fifty-two percent, average. Twelve percent were considered poor because their average scores fluctuated greatly, about thirty-one percent because they were easy, and about one percent because they were difficult.

Of the items on the Physics Examinations, three percent were found to have excellent discriminating power, while fifty-three percent were found to be average discriminators. Of the poor discriminators, nineteen percent were so categorized because their average scores fluctuated greatly, twenty-three percent because they were easy, and one percent because they were difficult.

Items in the following areas were classified as having poor discriminating power:

I. Biology
   1. "A plant embryo with a food supply and a protective coat is called (1) a fruit, (2) a seed, (3) an embryo sac, (4) an ovule." (January 1950, Part I, 4)

   2. "Tell whether each of the following is true or false and give your reasons." "Poison ivy can be destroyed by pouring

TABLE XIV

PERCENTAGES (APPROXIMATE) OF ITEMS OF EXCELLENT, AVERAGE AND POOR DISCRIMINATING POWER

Type of Test	Excellent Discriminating Power	Average Discriminating Power	Poor Discriminating Power
Biology	4%	60%	36%
Chemistry	1%	43%	56%
Earth Science	2%	53%	45%
Physics	3%	54%	43%

TABLE XV

THE POPULARITY OF THE ITEMS ON PARTS II OF THE SIXTEEN REGENTS EXAMINATIONS

Examination		Unpopular	Medium-Low	Average	Medium-High	Popular
Biology,	Jan. 1949	5		1,2,3,4,7	6,8	9
	June 1949			1,2,5,6,7	4	3,9
	Jan. 1950	8,9		1,2,5,6,7		3,4
	June 1950	6	1	5,7,8,9	3,4	2
Chemistry,	Jan. 1949	7	6	4,8		1,2,3,5
	June 1949	6	8	3,7		1,2,4,5
	Jan. 1950	8	7	6	5	1,2,3,4
	June 1950	7	8	4,6	3,5	1,2
Earth Science,	Jan. 1949			2,4,6	5,7	1,3,8
	June 1949		5	2,4,7	6	1,3,8
	Jan. 1950			4,5,6,7,8		1,2,3
	June 1950			1,3,7,8	4,6	2,5
Physics,	Jan. 1949	8		4,5,7	1	2,3,6
	June 1949		8	3,5,6,7	4	1,2
	Jan. 1950	7		6,8		1,2,3,4,5
	June 1950	7		8	1,4,6	2,3,5

salt water on its roots." (June 1949, Part II, 2 b_1) (See graph I)

II. Chemistry

1. "The reaction of the proposed hydrogen bomb involves a change of hydrogen to (1) argon, (2) radium, (3) helium, (4) uranium." (June 1950, Part I, 50)

2. "Describe how to make acetylene." (June 1950, Part II, 7 a)

III. Earth Science

1. "Feldspar may change to _____ when acted on by moist air." (January 1950, Part I, 9)

2. "Explain why relative humidity decreases as temperature increases." (January 1949, Part II, 2 d)

IV. Physics

1. "A balloon will rise until it displaces its own weight of air." (True-false) (June 1950, Part I, 31)

2. "The diagrams (of saxaphone and violin sounds) represent the wave patterns of the same note sounded on two different instruments. State one respect in which the sounds are similar. State one respect in which the two sounds are different." (January 1949, Part II, 7 d)

The following are examples of items showing excellent discriminating power:

I. Biology

1. "A tissue whose function is aided by the extensive branching of its cells is (1) blood, (2) epithelium, (3) nerve, (4) smooth muscle." (January 1949, Part I, 14) (See graph II)

II. Chemistry

1. "Isotopes of uranium have different (1) atomic numbers, (2) atomic weights, (3) numbers of planetary electrons, (4) numbers of protons." (June 1950, Part I, 48)

III. Earth Science

1. "The material deposited by a stream at the base of a mountain forms a (an)_____." (January 1950, Part I, 17)

IV. Physics

1. "A bottle can hold 120 grams of water. The same bottle can hold 96 grams of alcohol. The volume of the bottle is _____ cu. cm. The specific gravity of the alcohol is _____." (June 1949, Part I, 18, 19)

Summary

The following generalizations may be made relative to the discriminating power of the items:

1. Few of the items on any of the examinations could be considered as having excellent discriminating power.
2. The greatest percentage of the items were classified as average or poor discriminators.
3. There was an extremely small percentage of items showing consistently low average scores, while a large number of items had consistently high average or percentage scores. Of these latter, the majority were items on Part I.

Popularity of Items

Since a student has an opportunity to choose five out of eight or nine items on Part II of the examinations, it was decided to analyze the items with respect to their "popularity" with the students. To do this, the percentages of persons electing the various items were determined by dividing the number of students choosing an item by the total number of students obtaining a particular total score. This was done for each of the total score groups from (65) to 100. The percentages were then categorized on the basis of the following criteria:[10]

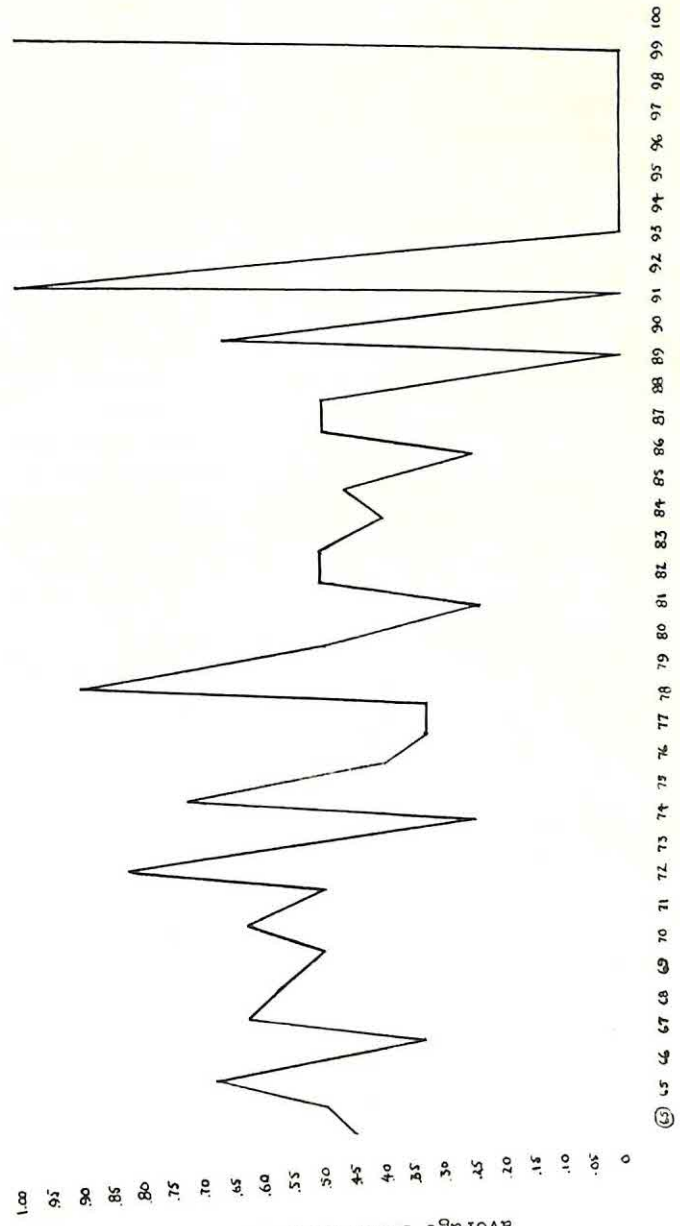
1. If the percentage was thirty or below, the item was considered to be "unpopular" with a single score group.
2. If the percentage was seventy or above, the item was considered to be "popular" with a single score group.
3. Items whose percentages ranged between thirty and seventy were considered to be of average popularity with a single score group.

Based on these criteria, the popularity of each item was tabulated for each total score from (65) to 100. These tabulations were then grouped as follows:

1. If the item was popular with seventy-five percent or more of the score groups, it was listed as popular.
2. If the item was unpopular with seventy-five percent or more of the score groups, it was listed as unpopular.
3. If the item was of average popularity with seventy-five percent or more of the scores groups, it was listed as being of average popularity.
4. If there were approximately equal numbers of items in both the unpopular and average

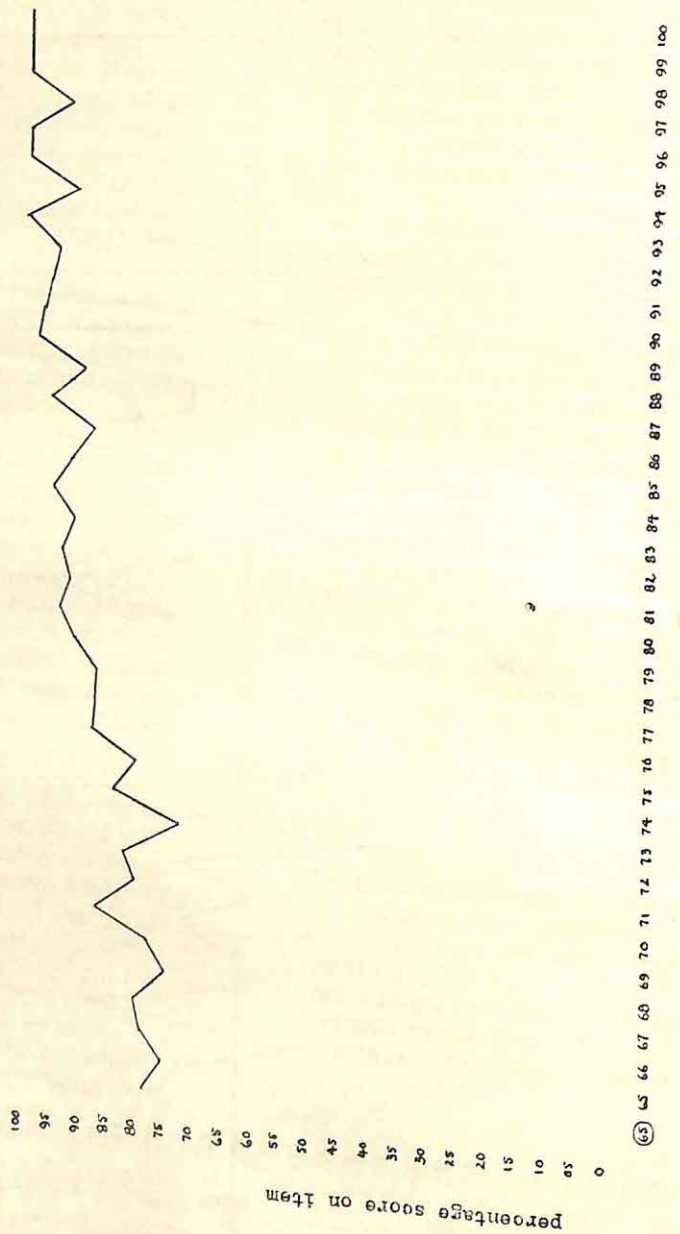
GRAPH I

AVERAGE SCORES OF ITEM 2 b_1, PART II. BIOLOGY EXAMINATION, JUNE 1949
(POOR DISCRIMINATION POWER)



GRAPH II

PERCENTAGE SCORES ON ITEM 14, PART I, BIOLOGY EXAMINATION, JANUARY 1949
(EXCELLENT DISCRIMINATION POWER)



categories, it was called an item of "medium-low popularity."

5. If there were approximately equal numbers of items in both the popular and average groups, it was considered to be of "medium-high popularity."

A listing of the items together with their respective classifications is found in Table XV.

An analysis of Table XV reveals that the majority (nineteen out of thirty-six) of the Biology items were considered to be of average popularity, while four were unpopular, and six popular.

On the Chemistry Examinations, the largest number of items, fourteen of thirty-two, were considered to be popular, four unpopular, and seven of average popularity.

It is interesting to note that there were no unpopular items on the Earth Science Examinations. Fifteen of thirty-two items were of average popularity, and eleven were popular.

Of the Physics items, thirteen of thirty-two were popular, ten were of average popularity, and three were unpopular.

The table reveals also that more of the low-numbered items are in the popular or medium-high categories; while the high-numbered items (those appearing at the end of the examinations) are more often unpopular. This is, of course, partly explained by the fact that many students (particularly those obtaining the higher total scores) answer the items in order—1, 2, 3, 4, 5—thus omitting those with higher numbers.

A further survey of the content of the popular and unpopular items indicates that, in general, the popular items are those concerned with the knowledge of factual information. The following are examples:

I. Biology

"In dogs, wire hair is dominant over smooth hair. A wire-haired dog is crossed with a smooth-haired dog. Show by keyed diagrams the cross which would result in: (1) a litter in which no smooth-haired pups could appear. (2) A litter in which a smooth-haired pup could be found." (June 1950, Part II, 2 a)

II. Chemistry

"Answer two of the following: (The atomic weights from the reference tables may be used to the nearest whole numbers, e.g., Cl = 35.457 becomes 35.) (a) How many grams of sodium hydroxide will be needed to neutralize 189 grams of nitric acid? (b) How many cubic feet of oxygen will be required for the complete combustion of 17 cubic feet of carbon monoxide? (c) How many liters of hydrogen sulfide gas will react with 99.3 grams of $Pb(NO_3)_2$?" (June 1950, Part II, 2)

III. Earth Science

"The following questions refer to the accompanying map: (a) Distinguish between contour line and contour interval. (b) State the contour interval of this map. (c) What is the highest possible elevation of hill A? How much higher or lower is hill B than hill A?..... (g) What does the map symbol at C represent?" (June 1949, Part II, 8)

IV. Physics

"A pulley system is used by a workman to raise a weight of 240 lb. a vertical distance of 24 feet. The workman's effort of 120 lb. moves through a distance of 72 feet. Find (1) the ideal mechanical advantage, (2) the actual mechanical advantage, (3) the efficiency of the pulley system." (January 1950, Part II, 1 a)

The unpopular items were found to be those involving the application of information, the use of elements of scientific method, the use of scientific attitudes, and the use of sciene in industry. The following are examples:

I. Biology

"A boy without a microscope wants to find out if there are bacteria on his fingers. (1) What is a culture medium and how is it sterilized? (2) List two important steps in his experiment following this sterilization. (3) What would indicate the probable presence of bacteria? (4) What evidence would he require to justify a conclusion that the bacteria had come only from his fingers?" (January 1950, Part II, 8 a)

II. Chemistry

"(a) Describe a process for making ethyl alcohol from molasses. (b) Name a by-product of this reaction. Give a use for the by-product. (c) Describe the manufacture of soap, mentioning the raw materials, the use of salt, and the by-product." (January 1949, Part II, 7 a, b, c)

III. Physics

"(a) An electric motor drives a d-c generator which is used to charge a lead storage battery. (1) State step by step three useful energy changes that occur, beginning with the input to the motor and ending with the energy in the battery. (b) Describe a laboratory experiment that may be used to illustrate two factors that affect the magnitude of an induced emf." (June 1950, Part II, 7 a_1, b)

In addition to the above analysis, still another survey was made regarding the relationship of popular and unpopular items with their

degrees of difficulty. It is interesting to note that all or part of sixty-seven percent of the unpopular or "medium-low" items were also considered as being difficult. Hence, as one would expect, it appears that students tend to avoid the more difficult items. Of the popular or "medium-high" items, sixty-four percent appeared among the easy.

Summary

A general review of the data concerning the popularity of items indicates the following:

1. The largest number of items were classified as being of average popularity, the second highest number, popular.
2. In general, the popular items appeared early in the examinations, while the unpopular items usually appeared near the end. This would seem to indicate that many students answered the items in order, thus omitting the last items.
3. As a rule, the unpopular items were the so-called "thought" items, or those requiring the application of facts and principles, or the use of elements of scientific method and scientific attitudes. The popular items were more often of the memory or factual type.

SECTION VIII

SUMMARY AND CONCLUSIONS

IN AN investigation as extensive as this one, any effort to prepare a detailed summary would be redundant. At the end of each section, the directors have listed a number of conclusions and implications, that apply to the respective section. Thus the summaries and conclusions that follow constitute little more than an epilogue, or perhaps a few philosophical observations.

1. This investigation, at least from the viewpoints expressed by the science teachers of New York State, fails to show that the Regents Examinations in Science are distasteful as many educators have implied. Most of the science teachers, with various types and degrees of reservations, seem to believe that the science program in New York State profits by their presence. A number of teachers do suggest specific changes, particularly that the examinations should place greater emphasis on measurement of general education objectives. Yet, such complaints apply to all educational programs and at all levels.
2. The phases of the investigation that dealt with the objective characteristics of the examinations indicate that they are far more reliable and valid than teacher-made tests and compare favorably with the commonly used standardized examinations in science. While the discriminatory power of the items on the various examinations did not prove to be high, those on standardized examinations fail to be much better.
3. In general, the examinations are not prejudicial to the interests of any particular group within New York State. While boys from the large high schools seemed to have the greatest achievement, and girls from small high schools, the least, the superiority and inferiority were neither consistent nor especially marked. Apparently the examinations appear to be as good (or bad) for one group as for another.
4. It does seem that a better system for scoring the examinations is indicated. Apparently teachers have been "on their own" more than may be considered desirable, and as a result a number of irregular scoring practices have occurred. Yet, none of these practices have been sufficiently widespread to cast doubt on the integrity of the science teachers of New York State as a whole.

It would seem, as a final statement, that this study failed to elicit the slightest bit of evidence that the examination system in New York State should be abolished. While it has revealed that the system of Regents Examinations in Science has weaknesses, the weaknesses are relatively the same as those that could be found with any mode of evaluation."

FOOTNOTES

1. Miller, David John and Mallinson, George Greisen. "An Investigation of the Attitudes of Teachers Toward the New York State Regents' Examinations in Science," Science Education, XXXVI (October 1952), 203-215.
2. Guilford, J. P. Fundamental Statistics in Psychology and Education (New York: McGraw-Hill Book Co., 1950), 161-164.
3. Peters, C. C. and Van Voorhis, W. R. Statistical Procedures and Their Mathematical Bases (New York: McGraw-Hill Book Co., 1940), 398.
4. Credit is due Professor Conway Sams, associate professor of mathematics at Western Michigan College of Education for helping to develop the statistical design, and for computing the corrections for the analysis of variance with unequal numbers of replications.
5. Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education (Boston: Houghton-Mifflin Co., 1953), 108-120.

6. Snedecor, George W. Statistical Methods (Ames, Iowa: Collegiate Press, 1946), 289-293.

7. The directors wish to express their gratitude to the many science teachers of New York State who contributed their time and efforts to evaluating the word lists.

8. Flesch, Rudolph. The Art of Plain Talk (New York: Harper and Brothers, 1946), vii + 210.

9. Buckingham, B. R. and Dolch, E. W. A Combined Word List (Boston: Ginn and Co., 1938), iii + 185.

10. Examinations. University of the State of New York, Division of Examinations and Testing, Albany, New York, December 1951.

11. It should be noted that since each student must select five of the eight or nine items, the chance factor would result in the selection of any item by approximately fifty-five percent. However, the cutting scores for this analysis of popularity have been set arbitrarily at thirty and seventy percent and hence make allowance for the chance factor.

A COMPARISON OF WECHSLER CHILDREN'S SCALE AND STANFORD-BINET SCORES FOR EIGHT- AND NINE-YEAR OLDS

FRANK C. ARNOLD
Bowling Green State University
WINIFRED K. WAGNER
Fremont, Ohio

THE HIGHLY verbal nature of the Stanford-Binet Intelligence Scale has long been considered a drawback in the psychological testing of certain children. The examiner in public schools must frequently use performance tests in cases where the Stanford-Binet seems inadequate because of its predominantly verbal character, e.g., the testing of children who work under a language handicap, those handicapped by speech and hearing defects, or those in whom the development of verbal and non-verbal abilities has been unequal (1).

The Wechsler Intelligence Scale for Children (7) or WISC, on the other hand, gives a verbal and performance score as well as a total score. The question arises, however, as to the degree of relationship between scores derived from this scale and those obtained with the Stanford-Binet, a scale already widely accepted. It is the purpose of this study to ascertain the relationship between these two scales for a sample of eight- and nine-year olds.

Several studies reported in the literature show a relatively high relationship between the two scales. Cohen and Collier (2), using first and second graders, report Pearson correlation coefficients between IQ's on the Stanford-Binet and WISC of .82 for the verbal scale, .80 for the performance scale, and .85 for the full scale. Frandsen and Higginson (3) found correlations of .71, .63, and .80 respectively between verbal, performance, and full-scale WISC IQ and Stanford-Binet IQ with unselected fourth graders. In a summary of four studies, Pastovic and Guthrie (5) report r's ranging from .63 to .83 between the Binet and the WISC Verbal Scale, from .57 to .75 for the performance scale, and .71 to .88 for the full scale. Krugman, Justman, Wrightstone, and Krugman (4) obtained correlations of the two scales for various age levels from 5 years, 6 months to 15 years, 6 months. They report r's ranging from .65 to .90 for the verbal scale, .60 to .75 for the performance scale, and .75 to .90 for the full scale. Findings similar to these are reported by Weider, Moller and Schramm (8).

Procedure

In the present study, a random sample of fifty children was selected from the eight- and nine-year olds in the third and fourth grades of the Bowling Green, Ohio, elementary school system. The Stanford Binet Intelligence Scale (Form L) and the Wechsler Intelligence Scale for children were then administered to each of the fifty subjects. All tests were administered by one person and were scheduled so that not more than one week elapsed between administration of the two scales. The Stanford-Binet administration preceded the Wechsler for one-half the subjects while the Wechsler administration preceded the Stanford-Binet for the remaining subjects. The order of administration was set on the basis of odd-even number of the subject.

Results

Means and standard deviations of the IQ's obtained by children in this sample on the two scales are presented in Table I. Comparison of these data with those reported by Terman and Merrill (6) and Wechsler (7) indicates that obtained results are quite similar.

In Table II* are presented the correlations between the Stanford-Binet and WISC scores with which we are concerned here. For purposes of comparison with reliability data of the Stanford-Binet, correlations between IQ's have been used as well as between mental ages and scaled scores. Correction of these r's has been made to take into account the differences between standard deviations obtained with this group and those reported by Wechsler (7).

In assessing the interchangeability of two scales for use in working with children, a logical approach would seem to be that of de-

*The correlation of .69 between the performance scale of the WISC and the Stanford-Binet is a correction of results reported in Winifred K. Wagner, A Comparison of Stanford-Binet Mental Ages and Scaled Scores of the Wechsler Intelligence Scale for Children for Fifty Bowling Green Pupils, unpublished Master's thesis, Bowling Green State University, 1951.

TABLE I

COMPARISON OF IQ's OBTAINED BY FIFTY CHILDREN ON STANFORD-BINET AND WECHSLER INTELLIGENCE SCALE FOR CHILDREN

Measure	Stanford-Binet	WISC		
		Verbal	Performance	Full
Mean	104.52	101.88	104.70	103.34
Standard Deviation	15.66	12.75	15.40	13.59

TABLE II

CORRELATIONS BETWEEN WECHSLER INTELLIGENCE SCALE FOR CHILDREN AND STANFORD-BINET

Item Correlated	WISC		
	Verbal	Performance	Full
Mental Age with Scaled Score	.77	.69	.81
IQ	.85	.75	.88
IQ (Corrected)	.88	.74	.90

termining whether the relationship between the two scales differs significantly from the reliability coefficient of one of the scales. Our concern here would be a comparison of the correlation coefficients obtained between the WISC and the Stanford-Binet and between Forms L and M of the Stanford-Binet. If the relationship found between the WISC and Binet is not significantly different from that between two forms of the Binet, then the use of the WISC as a substitute for the Binet in assessing IQ would seem reasonable. If, on the other hand, results do differ significantly, then other factors must certainly be considered in the substitution of one scale for the other.

The r of .93 given as the median value of relationship between Forms L and M of the Binet for ages six to sixteen was used as a basis for this comparison. The corrected correlation coefficients reported in Table II were transformed to Fisher z scores and differences computed between the z scores equivalent to these correlations and the z equivalent to a correlation of .93. From these differences, critical ratios were computed using the standard error of the difference between z's. The critical ratio found between the Binet reliability r of .93 and the corrected r of .88 obtained between the WISC Verbal Scale and the Stanford-Binet was 1.74, significant at the 10% level of confidence; for the corrected r of .74 between the WISC Performance Scale and the Binet was 4.37, significant at the .1% level of confidence; and for the corrected r of .90 between the WISC Full Scale IQ and the Binet was 1.15, significant below the 10% level of confidence.

Discussion

Data presented in Table I would seem to indicate that results obtained from this sample of children on the Stanford-Binet are quite similar to those reported for the standardization group (6).

Data concerned with relationship between the two scales is similar to that found by other investigators. Whether mental ages and scaled scores or IQ's are used, correlation coefficients are large enough to indicate a marked relationship between the two scales. This is particularly true for the verbal scale and full scale scores on the WISC. A squaring of the corrected r's given in Table II shows the common variance between the WISC and the Binet to be 77% for the verbal scale, and 81% for the full scale.

So far as this sample is concerned, the relationship between IQ's obtained for eight- and nine-year olds with the WISC and the Form L Binet is not significantly different from the relationship between IQ's obtained on Forms L and M of the Binet. So far as total score is concerned then, the WISC might very well be substituted for the Binet or the Binet for the WISC. From results of this study, the same would seem to be true for the WISC Verbal Scale. This would not seem to be true, however, for the WISC Performance Scale since the relationship found differs significantly from that between Forms L and M at the .1% level of confidence.

Clinically, it would seem that these findings have practical implications for the use of the various scales concerned. Total scores on the WISC or scores on the WISC Verbal Scale and the Binet would seem close enough to each other to offer practical interchangeability of the two scales. At the same time, the WISC Performance Scale would appear to be getting at a different facet of intelligence than is the total or verbal score of the WISC or the total score of the Binet. This study has not concerned itself with what these different scores may mean so far as prediction of various kinds of behavior is concerned. However, if broadened prediction is possible with the performance scale of the WISC while at the same time the total score and verbal score closely approximate that of a well-accepted tool, the WISC may prove to be a quite useful clinical instrument. Further research is necessary, of course, both to check findings of the present study and to determine the meaning of sub-scale scores of the WISC.

REFERENCES

1. Arthur, Grace. A Point Scale of Performance Tests, Clinical Manual (New York: The Commonwealth Fund, 1943).
2. Cohen, B. D. and Collier, Mary J. "A Note on the WISC and Other Tests of Children Six to Eight Years Old," Journal of Consulting Psychology, XVI (1952), pp. 226-227.
3. Frandsen, A. N. and Higginson, J. B. "The Stanford-Binet and the Wechsler Intelligence Scale for Children," Journal of Consulting Psychology, XV (1951), pp. 236-238.
4. Krugman, Judith I. and others. "Pupil Functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children," Journal of Consulting Psychology, XV (1951), pp. 475-483.
5. Pastovic, J. J. and Guthrie, G. M. "Some Evidence on the Validity of the WISC," Journal of Consulting Psychology, XV (1951), pp. 385-386.
6. Terman, L. M. and Merrill, Maude A. Measuring Intelligence (New York: Houghton-Mifflin, 1937).
7. Wechsler, D. Wechsler Intelligence Scale for Children, Manual (New York: The Psychological Corporation, 1949).

8. Weider, A. and others. "The Wechsler Intelligence Scale for Children and the Revised Stanford-Binet," Journal of Consulting Psychology, XV (1951), pp. 330-333.

ERRATA

We regret that the following three tables were inadvertently left out of author Evan R. Keislar's article "Peer Group Rating of High School Pupils with High and Low School Marks," published in the June 1955 Journal of Experimental Education.

TABLE I

CORRELATIONS OF EACH OF TWELVE TRAIT RATINGS WITH OTIS I.Q. AND SCHOOL MARKS FOR 126 BOYS AND 128 GIRLS

Trait	Otis I.Q.		School Marks	
	Boys	Girls	Boys	Girls
1. Talkative - silent	-.02	-.06	.10	-.22
2. Old acting - young acting	.10	.17	.21	.06
3. Friendly - unfriendly	.03	.14	.09	.24*
4. Likes schoolwork - dislikes schoolwork	.44*	.38*	.75*	.75*
5. Considerate - inconsiderate	.17	.09	.36*	.26*
6. Popular - unpopular (with opposite sex)	-.12	.13	-.07	-.10
7. Persistent - not persistent	.38*	.23*	.63*	.49*
8. Welcomed - ignored (by same sex)	-.09	.18	.03	.17
9. Puts studies first - puts studies last	.30*	.36*	.70*	.70*
10. Conceited - not conceited	-.16	-.04	-.27*	-.22
11. Cheerful - sad	-.02	.08	.05	.09
12. Boys athletically competent - incompetent	-.22		-.07	
12. Girls influential - not influential		.20		.31*

*Significantly different from zero at the .01 level.

TABLE II

DIFFERENCES ON TRAIT RATINGS BETWEEN TWO GROUPS OF HIGH SCHOOL GIRLS MATCHED FOR OTIS I. Q. BUT DIFFERING IN SCHOOL MARKS

Trait	School Marks	Based on 27 girls in each group: Mean	σ	D	s_D	t	Level of Signif.
Talkative - Silent	Low	57.3	9.0	7.8	3.3	2.39	.05
	High	49.5	12.9				
Old acting - Young acting	Low	51.9	6.4	1.5	1.8	.87	...
	High	50.4	6.3				
Friendly - Unfriendly	Low	53.0	7.0				
	High	57.5	7.8	4.5	1.9	2.39	.05
Likes schoolwork - Dislikes schoolwork	Low	39.9	5.6				
	High	58.0	8.3	18.1	1.8	9.87	.001
Considerate - Inconsiderate	Low	49.4	6.7				
	High	53.0	5.2	3.6	1.6	2.24	.05
Popular - Unpopular (with opposite sex)	Low	55.0	9.9	6.2	2.2	2.84	.01
	High	48.8	9.9				
Persistent - Not persistent	Low	46.8	5.0				
	High	54.9	5.7	8.2	1.2	6.66	.001
Welcomed - Ignored (by same sex)	Low	52.3	4.1				
	High	54.0	6.2	1.7	1.4	1.23	...
Puts studies first - Puts studies last	Low	44.4	4.4				
	High	54.4	4.6	10.0	.94	10.62	.001
Conceited - Not conceited	Low	51.0	6.3	3.7	1.7	2.18	.05
	High	47.3	5.3				
Cheerful - Sad	Low	53.2	5.8				
	High	54.2	6.0	1.0	1.6	.61	...
Influential - Not influential	Low	48.8	4.5				
	High	53.4	5.9	4.6	1.2	3.79	.001

Note: All figures reported have been rounded off to one decimal place except for the values of t.

TABLE III

DIFFERENCES ON TRAIT RATINGS BETWEEN TWO GROUPS OF HIGH SCHOOL BOYS MATCHED FOR OTIS I. Q. BUT DIFFERING IN SCHOOL MARKS

Trait	School Marks	Based on 35 boys in each group: Mean	σ	D	s_D	t	Level of Signif.
Talkative - Silent	Low	54.0	11.9	3.1	2.9	1.07	...
	High	50.8	11.6				
Old acting - Young acting	Low	46.2	8.0	4.6	2.1	2.25	.05
	High	50.8	8.1				
Friendly - Unfriendly	Low	50.2	6.6	2.2	1.7	1.33	...
	High	52.4	7.2				
Likes schoolwork - Dislikes schoolwork	Low	43.5	10.3	13.8	2.2	6.31	.001*
	High	57.3	9.2				
Considerate - Inconsiderate	Low	47.8	5.6	4.1	1.3	3.11	.01
	High	51.9	5.5				
Popular - Unpopular (with opposite sex)	Low	48.8	8.1	.3	2.2	.13	...
	High	49.1	8.1				
Persistent - Not persistent	Low	46.6	4.7	5.4	.9	5.91	.001
	High	52.0	4.9				
Welcomed - Ignored (by same sex)	Low	49.4	6.7	3.4	1.7	1.98	...
	High	52.8	6.8				
Puts studies first - Puts studies last	Low	45.0	5.7	9.3	1.4	6.47	.001
	High	54.3	5.8				
Conceited - Not conceited	Low	51.5	5.4	3.0	1.3	2.36	.05**
	High	48.5	5.1				
Cheerful - Sad	Low	51.8	4.5	1.6	1.0	.63	...
	High	52.4	5.4				
Athletically competent - Athletically incompetent	Low	47.7	6.7	3.5	2.0	1.71	...
	High	51.2	9.9				

Note: All figures reported have been rounded off to one decimal place except for the values of t.

* For the distribution of trait scores the hypothesis of normality could be rejected at the .02 level but not at the .01 level.

**For the distribution of scores the hypothesis of normality could be rejected at the .05 level but not at the .02 level.

Journal of Experimental Education

Volume XXIV December, 1955 Number 2

THE EFFECTS OF A "CAUSAL" TEACHER-TRAINING PROGRAM AND CERTAIN CURRICULAR CHANGES ON GRADE SCHOOL CHILDREN*

RALPH H. OJEMANN, EUGENE E. LEVITT
WILLIAM H. LYLE, Jr., MAXINE F. WHITESIDE
Child Welfare Research Station
State University of Iowa

THE PURPOSE of this paper is to report the results of a learning program designed to help the child develop a "causal" orientation toward his social environment. The learning program used in this study involved both the training of teachers and the use of certain special curricular content.

The meaning of the term "causal" as used here has been detailed in an earlier publication (4). Briefly, it recognizes that human behavior is produced by many factors and that one can distinguish between an approach to a given behavior incident which recognizes and takes into account the variety of factors that may have produced it as compared with an approach that considers mainly the overt form of the behavior.

The use of the term "causally oriented curricular content" arises from the discovery that present curricular content relating to human behavior as found, for example, in current social studies readers and texts is largely non-causally or surface oriented (7, 8).

The importance of specifying the learning program as involving the training of teachers rests on the following: Previous data have suggested that teacher behavior toward children is essentially non-causally oriented. Since our culture is for the most part likewise oriented and since teachers have come up through that culture, this situation is not unexpected. But the tendency toward a non-causal orientation in the daily behavior of the teacher becomes important when we consider the problem of developing a causal orientation in the child. This may be explained as follows:

When we teach arithmetic we can conceive of a situation in which the teacher would teach the child to perform the various number operations accurately while at the same time he (the teacher) would make a number of "mistakes" on his income tax report. The child need never see these "mistakes" and thus they would not directly influence his learning.

But in teaching an approach to human behavior the situation is different. The teacher must of necessity interact with the pupil. Through the approaches he makes to the pupil he provides a demonstration from which the pupil learns. If he approaches the pupil in a non-causal way, the pupil is experiencing a demonstration of a non-causal approach.

Thus, in the area of human behavior the teacher teaches in two ways: He teaches through the content studied and through the daily demonstrations he provides. In a previous study (10) evidence was obtained indicating that it is difficult to develop a causal orientation if the regular classroom teacher and content remain essentially non-causally oriented and causal content is introduced for, say, one period a day by a trained but "imported" teacher.

Testing the effects of a learning program using trained teachers and causally oriented curricular content involved: (a) a training program for teachers, (b) a plan for changing curricular content, (c) an appropriate experimental design, and (d) the gathering of data and analyses of results. Each item will be briefly described in turn.

The subjects of this study were four teachers and their pupils, each classroom matched with two control groups. One of the teachers was from the fourth grade, one from the fifth and two from the sixth grade. All were from the school system of a midwest industrial town of about 75,000 population. Since this investigation is part of a long range program, it was desired to develop a group of trained teachers who gave promise of

*The preparation of this paper was supported by Research Grant MH-301 from the National Institute of Mental Health, U. S. Public Health Service.

remaining in the system for several years. Accordingly teachers were selected on this basis by the school administration in consultation with the investigators. Data relative to the experimental subjects will be presented in a description of the experimental design.

Teacher Training Program

Our plans involved providing teachers with one month of intensive work during the summer and following through with group conferences every three weeks during the school year. These conferences were intended to give the teachers opportunity to discuss any questions or problems which might arise during the year as closely as possible to the time they might arise.

The month's program of intensive work was set up under circumstances similar to the usual academic situation. Limited credit on a minor problems basis was allowed for those teachers who indicated that they wished to receive academic credit. The program was organized in terms of six units, all but one to be completed during the four week period. The description of the units, the time devoted to each, and the reason for their inclusion in the program are presented below:

Unit 1. Developmental Problems of the Normal Child—three hours per week. The primary purpose of this unit was to draw the teachers' attention to the fact that "having problems" does not necessarily make a child a problem child. Emphasis was upon the kinds of developmental tasks children face at various ages, the kinds of basic learnings which are necessary for proper handling of these tasks, and the problems which are created when tasks appropriate to a particular age level are not learned before the following level. Selected portions were assigned of F. Redl and W. W. Wattenberg, Mental Hygiene in Teaching; R. J. Havighurst, Human Education and Development, Association for Supervision and Curriculum Development Yearbook, 1950, Fostering Mental Health in Our Schools; and Gladys Jenkins et. al., These Are Your Children. It was our intention for teachers to understand from these materials that children are continually facing problems and that problems are a necessary result of the child's expanding social environment. Instructors: S. L. Zelen and C. D. Smock.

Unit 2. Personal Problems of Everyday Life—five hours per week. This unit was set up, but not labeled, as 20 one-hour sessions of group psychotherapy. It was presented to the teachers as an opportunity to "extend the individual's understanding of the problems people ordinarily meet in their daily living; to acquaint the individual with general psychological principles which have maximum relevance to these problems both with regard to handling personal problems which exist currently and the off-setting of present behavior trends which could conceivably lead to future problems; and to assist in the development of personal techniques for meeting the frustrations which most of us normally encounter." The preparation of an extensive personal autobiography was required following essentially the lines presented in Stogdill's Mental Hygiene workbook entitled Objective Personalty Study. Extensive comments were made about the material contained in the units of the workbook which were intended to stimulate thinking about their personal experiences. The teachers were encouraged to explore the extent to which their own personal biases and predilections might structure the classroom situation in the hope that this might minimize the extent of influence of that bias. No individual sessions were held with members of the group except when they presented themselves to ask for individual discussion, at which time they were encouraged to raise their questions for discussion in the group. That is, an explicit attempt was made to focus attention on the group situation and to bring discussion material to the group rather than to take material away for individual sessions. Members of the group were assigned collateral reading from Philip Eisenberg, Why We Act As We Do, and from Hugh Cabot and Joseph A. Kahl, Human Relations, Volume I, Concepts. These readings plus the autobiographical material provided the vehicle for group discussion. Instructor: W. H. Lyle.

Unit 3. Action Research in the Classroom—two hours per week. An essential part of this unit was an attempt to discourage the teacher from having too much confidence in her observations and her ability to predict from them. Data were presented on the problems involved in the determination of the reliability of observational techniques and the predictive efficiency of these observations. Some methods for placing her own observations in a research framework were presented and the teachers were encouraged to make their observations in a somewhat more systematic manner. Selected papers were used as source material and no outside reading was assigned. Instructor: E. E. Levitt.

Unit 4. The Causal Approach to an Understanding of Human Behavior—two hours per week. The primary purpose of this unit was to acquaint the teachers with the background of the

project, its origin, and its present status. An additional function of this unit was to acquaint them with the special materials which had been developed by the project. Instructor: R. H. Ojemann.

Unit 5. Meeting Classroom Problems—three hours per week. This unit was under the direction of an experienced classroom teacher who had been working with the project for the past two years and had had direct experience in classroom situations. This was a technique-centered unit. That is, our attempt was to help the teacher to utilize known techniques in the handling of classroom problems and to develop special techniques which would allow her to meet the daily problems arising in the classroom. "Typical" classroom situations were presented to the teachers to give them some experience in understanding what would be surface ways of handling these situations as opposed to possible causal methods. The probable effects of the methods were compared. Our concern was with individual needs, but it was our feeling that most would be accomplished if group and individual needs were met jointly. Many previous attempts to encourage the teacher to take individual needs of children into consideration have failed to consider that this can only be accomplished, or at least accomplished most effectively, within the framework of good group control. In this manner the constructive forces of the group are at the disposal of the teacher in meeting individual problems. All of the materials developed and used by the project previously were discussed with the teachers. In a sense, this unit might be considered as a practicum companion for Unit 1, since assistance in the handling of developmental tasks formed an important part of this unit. Instructor: Mrs. Maxine Whiteside.

Unit 6. Practicum in the Preparation of Special Materials—two hours per week. This unit represents our attempt to insure two-way communication. The project personnel felt the need of assistance in the adaptation of materials to be used. It was our belief that those individuals closer to the teaching situation might adapt materials to the classroom situation more effectively both from the point of view of interest and appropriateness. The participating teachers were encouraged to write materials to replace those we had developed, to extend such materials, and to develop new materials utilizing the strengths in their own professional background. All materials were discussed with the project member acting as advisor and the joint suggestions incorporated. For the most part, this proved to be a continuing project on which the teachers worked during the entire year. Instructors: Staff.

Conferences with Experimental Teachers During School Year

Twelve meetings were scheduled during the school year or approximately one meeting every three weeks. The general purpose of these meetings was to provide an opportunity for the teachers to ask questions concerning the classroom work they were doing. It was recognized that the actual practicing of the causal approach in the classroom would give rise to more specific questions which could not be fully anticipated during the summer training program. In addition, the meetings furnished the staff with an opportunity to discuss additional topics with the teachers.

One or more members of the staff led discussions on various topics which can be grouped under seven general headings.

Materials—At each meeting the teachers were given an opportunity to ask any questions about the materials they were using. At six of the meetings, questions were presented and discussed. Other meetings were used for extending teachers' background in child behavior and discussions relative to practicing the causal approach in the classroom.

At one meeting toward the close of the program the teachers were asked to suggest, on the basis of their experience, the teaching sequence for using the materials. A discussion of the merits of teaching one type of material before another and the like, resulted in agreement as to the most useful sequence according to their classroom organization.

Pupils—One of the main topics of the first meeting was a discussion of specific classroom situations which the teachers had faced, comments on the surface and causal methods to handle such situations plus a description of the way the teachers had handled the situations. Part of nine other meetings was spent discussing this topic. In this way, the teachers were given an opportunity to check their own behavior as surface or causal as well as the behavior of the pupils.

For example, one teacher had been observing a girl who seemed to play with no one, who stayed by herself but had made it known that she wanted to associate with others. Meetings with the parent, conversations with the pupil were described after which teachers asked questions to obtain additional information, such as the teacher's hypothesis concerning the causes of the described behavior. The group then made and evaluated recommendations for possible methods of dealing with this situation.

Records—At the seventh and twelfth meetings time was devoted to discussing what information the teachers would like to have about their pupils in order to better practice the causal approach in the classroom.

Additional background in child behavior—The teachers were given an opportunity to question members of the staff relative to the findings of investigations of a variety of behavior patterns. The teachers' questions arose primarily from observations of pupils in their rooms which prompted them to inquire about studies which might further their understanding of the pupils. Though some background had been provided during the summer program, a re-presentation was advantageous because of the teachers' actual observation of the behavior being discussed.

For example, one question was "Would you discuss the 'shy child' in general and then consider a specific case which I will describe?"

Outcomes—At the second meeting the teachers were asked to assist in the preparation of a "Tentative List of Outcomes Which Might Be Expected as a Result of Teaching the Causal Approach." After a tentative list had been prepared they were asked to refer to it often during the school year and then toward the end of the second semester, select the outcomes which they felt might be a result of teaching the causal approach at their particular grade level. The purpose of this exercise was to utilize the teachers' experiences in making a tentative estimation as to what aspects of the causal approach may be developed at the respective grade levels.

Evaluation—Toward the middle of the year the teachers were asked to evaluate the training program of the previous summer by answering the following questions:

1. What do you feel were the most valuable parts of the training program last summer? Please list at least two or three, with comments.
2. If a new group of teachers were to be trained, what changes in the training program would you suggest?

Results—The last meeting was primarily concerned with the presentation of the statistical analysis of the results of the program.

Development of Curricular Content

As indicated above, previous studies had demonstrated that content dealing with human behavior as currently found in elementary readers, social studies and health texts is essentially surface or non-causal in nature. It was, therefore, necessary to develop more causally oriented content. To accomplish this a variety of materials were prepared. Some of these materials were available from previous studies, some were prepared during the course of this investigation.

In describing the preparation of causal content a statement of the concepts and appreciations which constitute the goals of the learning program may facilitate discussion. We wish to help the child to understand and appreciate more about how his social environment operates. He is taught that there are many ways in which a given behavior pattern may develop, that causes are complex, that people are faced with many different situations which they are trying to work out, that they use a variety of methods for this, that additional methods may be available and that all the methods may be considered in terms of the effects they have.

In contrast to such concepts as these, children under present usual conditions are taught essentially what people do and primarily a judgmental approach to the behavior without first seeking an understanding of how it came about.

The situation appears somewhat comparable to that which prevailed in man's reaction to his physical environment. At one time man took a more or less arbitrary approach to his physical environment. It is only relatively recently when we consider the span of human history that he learned a more dynamic approach.

A list of some of the elementary concepts represented in the causal orientation has also been reported elsewhere (4).

The nature of the curricular content is further revealed by the specific materials developed. It is possible here only to list the various types. Readers who are interested in examining the materials at first hand may obtain copies from the investigators.

The types of materials are as follows:

1. Introduction to the causal approach by the story method—the "Teachers Manual for Behavior Materials in the Primary Grades" is a collection of twenty-seven stories grouped in sections for use at different grade levels. Each story deals with a particular behavior pattern. Preceding each story the manual supplies some background for the teachers. These materials have been described in earlier publications (7).

The story is introduced and read by or to the pupils and is followed by a discussion designed to guide the pupils into thinking of the "reasons for the behavior" which were described in the story. The teacher keeps two general questions in mind during discussions:

1) Did the children understand the differences between thinking of causes and not thinking of causes?
2) Did the children gain ideas of ways to meet ordinary problems so as to help each participant grow?

Stories for use in the intermediate grades were also written with a broader scope than those for use in the primary grades.

2. Expository presentation of causal approach

as it applies both to development of behavior and the consideration of the effects of behavior—two pamphlets bearing the titles "Two Ways to Look at How People Act" and "When We have to Decide" provide in expository form the differences between the surface and causal approaches.

3. A series of workbooks which served as introductory units to social studies and health:

Book I: How considering causes affects our reaction to behavior

Books II and III: How people work out feelings of self-respect and "counting-for-something"

Books IV and V: How physical differences, experiences and opportunities may affect different people

Book VI: How past experiences affect methods people use

The booklets provided a variety of exercises to be written out, unfinished situations for which endings were to be written or role-played and the like.

4. Revised units in history and geography—sections of history and geography were revised to incorporate the elementary principles of human behavior. For example, the unit on "The South" was revised to include discussions of how geographical and cultural conditions may influence the situations people face and the methods they employ to work them out.

5. Units on the use of the room council—the material prepared by Stiles (9) for helping pupils apply the causal approach in room council discussions has been described in previous publications.

In the preparation of these materials, considerations were given to pupil interest, pupil experiences and vocabulary burden. The Dolch, Buckingham "Combined Word List" and Green's "Iowa Spelling Scale" aided in checking vocabulary in material to be read by pupils. Listening vocabulary was scaled higher than reading vocabulary in recognition of the differences between the two.

Relating new situations or experiences to familiar ones is a technique often used by teachers. This practice was taken into consideration in the writing of materials with one precaution. As is explained in the teachers' manual of primary materials: "Since every child is engaged in working out his own problems, it was felt that if the material dealt only with school and community situations of children like themselves, they may become so engrossed in their immediate problems that they miss the larger more objective appreciation. Accordingly, situations involving children older and younger than themselves, and children from quite different environments as well as some situations involving children like themselves are included."

Since the incorporation of the causal approach in teaching materials is relatively new, readers who are interested in detail are encouraged to examine the original materials. Particular questions vary with the background of the reader and it is not possible to anticipate all of them. As a guiding principle it may be helpful to keep in mind that the purpose of the learning program is to help the child gain more appreciation how his social environment operates just as physical science attempts to build an appreciation of how the physical environment operates.

Experimental Design and Analysis of Results

The evaluation of the teacher training program was actually concerned with pupil development rather than with teacher development per se. There were two reasons for this approach: a) the primary motive for the training of the teachers was to affect the pupils in certain ways, b) the number of teachers was obviously too small to permit any reliable measurement of teacher characteristics directly. The evaluation procedure is described in detail in the following sections.

Control teachers—Two control teachers were selected for each of the four experimental group teachers. The control teachers were matched with the respective experimental teacher, insofar as it was feasible, on a number of dimensions which might affect experimental results. These variables were age, sex, number of years of teaching experience, and educational level. The data are shown in Table I. The control teachers were selected from the same school system. The twelve teachers represented ten different elementary schools.

It would have been desirable to have been able to control other potentially pertinent factors like teaching ability. However, an analysis of the available literature indicates that such expressions as teaching ability are rather nebulous and not easily defined or measured. It seemed preferable to deal with concrete measures and to assume that meaningful uncontrolled variables were randomly distributed among the groups of teachers.

In addition to their training, the experimental teachers had been provided with various materials for use in teaching the "causal approach" in the classroom. The purpose of the double control group was to attempt to determine the effectiveness of these materials alone. Toward this end, the teachers in $Control_1$ were invited to secure and make use of such of the materials as they wished. The purposes and modes of use of the materials were outlined briefly. Their use was not, however, required of the $Control_1$ teachers and no attempt was made to insure that they were used. A check on the kinds of

TABLE I

COMPARATIVE BACKGROUND DATA OF EXPERIMENTAL AND CONTROL TEACHERS

Teacher	Age	Sex	Years Teaching	Educational Level
		Fourth Grade		
Experimental	26	F	5	B. A.
$Control_1$	27	F	5	B. A.
$Control_2$	26	F	4	B. A.
		Fifth Grade		
Experimental	52	F	36	B. A.
$Control_1$	50	F	30	B. A.
$Control_2$	52	F	32	B. A.
		Sixth Grade (I)		
Experimental	44	F	23	B. A.
$Control_1$	40	F	17	B. A.
$Control_2$	50	F	28	B. A.
		Sixth Grade (II)		
Experimental	26	F	5	B. A.
$Control_1$	28	F	5	B. A.
$Control_2$	26	F	4	B. A.

TABLE II

MEAN IQ SCORES BY CLASS

	Fourth Grade	Fifth Grade	Sixth Grade (I)	Sixth Grade (II)	Total
Experimental	110.89	108.47	111.74	105.68	109.20
$Control_1$	109.76	110.08	101.84	101.20	105.72
$Control_2$	106.31	109.94	108.63	107.88	108.19
Total	109.53	109.53	106.78	104.40	107.48

material actually used and the number of hours devoted to them was made at the conclusion of the investigation (see page 110).

All eight of the control teachers expressed a desire to be in $Control_1$, i.e., were apparently interested in the materials. Teachers were assigned at random to the two control groups. The untreated control group, i.e., the one in which the teachers had no contact whatsoever with the experimenters, is designated as $Control_2$.

The pupils—The matching of teachers need not, of course, have any affect on the disposition of pupils within the various groups of classes. It was necessary to be reasonably certain that the pupils in one or another of the groups were not superior in any relevant way. Age and sex were controlled automatically by the methods of assignment of pupils to classes in the school system. Some discrepancies in the sex ratio might occur, but earlier work with the tests to be used have revealed no systematic sex differences in performance.

Intelligence is likely to be an important factor, as it usually is in studies of this type. IQ scores on the Otis Self-Administering Test, Intermediate form, were secured from school records for all pupils who participated in the testing program. The mean IQ scores by class are shown in Table II. The results of a treatment-by-grades analysis of the variance of IQ scores are shown in Table III.

The analysis of variance[1]* of IQ scores yields entirely negative results. There are hence no significant differences in intelligence either among the three treatment groups, or among grades, or among individual classes. We shall not, therefore, be able to attribute differences in performance to intelligence.

We have been able to control age, sex, and intelligence among the pupils in the various classes. It is entirely possible that there are other, unknown variables which are pertinent to the experimental design, a not uncommon occurrence. Again we shall assume that the pupils are randomly distributed among the groups with respect to such variables.

The tests—Two instruments were used in the evaluation proper. The first of these, the Problem Situations Test (PST), has been the subject of considerable investigation. Its development is described elsewhere (5). The PST is a 22-item multiple-choice test in which the subject is faced with a number of instances of misbehaviors or deficiencies of children and is required to deal with them either from the point of view of an authority figure or from his own point of view. There are six possible responses for each situation, three punitive and three non-punitive. The punitive responses prescribe verbal or physical punishment, deprivation, or coercion. The responses were obtained from an open-end form of the test administered earlier to a group of fifth grade children.

The PST is considered to be a measure of punitiveness in the child, that is, his willingness to be immediately punitive in a hypothetical situation where no retaliation is anticipated. The score for punitiveness is the number of punitive responses to the 22 situations. The PST has been shown to be related to authoritarianism and parental disciplinary methods (6) and to extrapunitiveness and intrapunitiveness as measured by the Rosenzweig Picture-Frustration Study (3). The reliability of the PST has been estimated as .77 using the Kuder-Richardson formula 20, based on data obtained from the earlier studies. In a sample of fifth grade pupils the correlation between the PST and IQ was found to be -.29 (6) which indicates that only a small fraction of the variance of the test is due to intelligence.

The second instrument was the Causal Test (CT). The CT has not yet been widely investigated, though it appears to have considerable promise. It is a 30-item true-false type, the individual items being based on eight descriptions of behavior. The test attempts to tap the child's awareness of the dynamic, complex, variable nature of human motivation, though it does not require that he have any specific knowledge of the causes of behavior themselves. This awareness and its hypothesized behavioral concomitants has been called "causality" (4). The test is scored inversely, i.e., for non-causality, so that the higher the score, the less causal the subject. This was done so that the CT would vary directly with the PST. The CT has been found to correlate -.36 with intelligence in fifth grade pupils and to have a Kuder-Richardson reliability of .63. The latter is rather low, but it should be borne in mind that attitude and personality tests with young children cannot be expected to have reliabilities of such measures with adults.

A more detailed description of the CT will be found in another forthcoming publication (2).

Experimental procedure—The tests were administered to all twelve classes on September 29 and 30, 1954, approximately three weeks after the beginning of the fall semester. (The formal body of the teacher training program had ended in July, 1954.) This administration will be referred to as the pre-testing. The second administration, or post-test, took place on April 12 and 13, 1955, approximately six and one-half months later. The tests were administered by three regular project staff members, each of whom tested the same four classes in September and April. No administrator tested more than two classes in any one treatment group. The

*All footnotes will be found at the end of this article.

TABLE III

ANALYSIS OF VARIANCE OF IQ SCORES

Source	d.f.	SS	MS	F	P
Treatments	2	565.95	282.975	1.762	> .10
Grades	3	1028.78	342.927	2.135	.10
T × G	6	1273.21	212.202	1.321	> .20
Within Cells	228	36621.96	160.623	.....	...
Total	239	39489.90	.....	.....	...

TABLE IV

LOSS OF SUBJECTS DUE TO EQUATING CLASS N's OVER TREATMENT GROUPS

Grade	Experimental	$Control_1$	$Control_2$	
		Problem Situations Test		
Fourth	19 (-0)	23 (-1)	17 (-1)	
Fifth	21 (-2)	22 (-0)	16 (-0)	
Sixth (I)	20 (-1)	31 (-9)	24 (-8)	
Sixth (II)	24 (-5)	28 (-6)	24 (-8)	
Total Eliminated	8	16	17	41
N Remaining Per Class	19	22	16	
		Causal Test		
Fourth	19 (-0)	25 (-0)	16 (-0)	
Fifth	26 (-7)	26 (-1)	17 (-1)	
Sixth (I)	20 (-1)	31 (-6)	23 (-7)	
Sixth (II)	23 (-4)	28 (-3)	25 (-9)	
Total Eliminated	12	10	17	39
N Remaining Per Class	19	25	16	

tests were administered in the same order and no time limits were set. Despite this leniency, there were a number of incomplete protocols in every class, especially for the pre-testing. These were invariably discarded.

The number of subjects who successfully completed both pre- and post-tests varied from class to class for both of the measures. In order to avoid complicating an already complex statistical analysis, it was necessary to equate the numbers of subjects either over the treatment groups or over the grade levels. The former was the technique chosen since it involved the smaller loss of subjects for both tests. All subjects were first listed randomly, then the required number was eliminated according to a table of random numbers. Table IV shows the original Ns, the number eliminated, and the remaining N for each class and test.

The elimination of subjects changed the mean scores per class only slightly, which is the anticipated result when subjects are randomly rejected. For the comparison of pre- and post-test results there are 19 subjects in each experimental class, 16 subjects in each $Control_2$ class, and 22 subjects in $Control_1$ classes for PST, and 25 in $Control_1$ classes for the CT. The total number of subjects will be 228 for the PST and 240 for the CT.

If the teacher training program has had the desired results, we would expect that the pupils taught by the experimental teachers would show greater reductions in PST and CT scores than the pupils taught by the control teachers. If the use of learning materials alone has any significant effect, we would also expect the $Control_1$ classes to improve more than the $Control_2$ classes, although, of course, this difference is not basic to the evaluation of the training program.

Statistical analysis—In a design of this type we would expect to find some random (though insignificant) differences in pre-test scores among the treatment groups. Since these pre-test differences may have some effect on the post-test scores, it would be desirable to eliminate them by means of some statistical technique. Hence the appropriate statistical procedure is an analysis of covariance.

If the data do not justify the assumptions necessary for the application of a covariance analysis,[2] there remain two alternate analyses. The first of these is simply to accept the post-test scores as a valid index of treatment effects on the assumption that the lack of significance of differences among pre-test scores means that the groups were actually equated prior to treatment. The second is a sign test (1) based on preminus post-test scores, a non-parametric method requiring no assumptions.

Results - The PST: Pre-Test—The mean pre-test scores on the PST are shown in Table V.

Obviously there are arithmetic differences between class means, although the differences between mean scores for the three treatments, 5.17 for the experimental group, 5.86 for $Control_1$ and 5.38 for $Control_2$ are quite close together. The results of an analysis of variance of these scores are shown in Table VI.

The analysis reveals no significant difference between the treatment means ($F = 0.580$) and no significant interaction ($F = 1.915$). Differences between grades are significant ($F = 3.516$, $P = <.02>.01$) but this is of no consequence for the experimental design. The absence of differences between treatment means and the lack of interaction indicate that random sampling has been accomplished. That is, the classes have been assigned at random to the treatment groups and are thus well matched. We may conclude that this phase of the testing with the PST has been successful.

The PST: Post-Test—The post-test means are shown in Table VII. The pre-test means of Table V are included for comparative purposes.

The experimental group, with a pre-test mean of 5.17, dropped to 2.39 on the post-test. $Control_1$ fell from 5.86 to 5.14, a change of less than three-quarters of a point. $Control_2$ rose slightly, from 5.38 to 5.67. The experimental classes show a unanimous decrease in mean score, the smallest decrease, that for the fourth grade, being over 1.25 points. Three of the four classes in $Control_1$ show decrements, although the overall decrease is much less than that for the experimental group. Two of the $Control_2$ classes show increases and two show decreases, the net being an increase of 0.29 points.

We now proceed to an analysis of variance of the post-test scores, which is shown in Table VIII.

We find on post-test that the difference between treatment means is now highly significant ($F = 23.4$, $P = <.001$). The differences by grades remain significant, though of no consequence. The interaction also remains insignificant.

Before we can proceed to adjust the post-test scores by covariance, it is necessary to test for homogeneity of regression, a key assumption in the application of covariance. For this test we break down the adjusted within cells sum of squares for the post-test scores into two components, the sum of squares for differences among group regression lines and the sum of squares for deviations from the group regression. The mean square for the former divided by the mean square for the latter constitutes the F-ratio for the test of homogeneity of regression. The degrees of freedom are the number of regressions minus one for the numerator and N minus twice the number of regressions for the denominator.

For the PST, the MS for differences among

TABLE V

MEAN PRE-TEST SCORES ON THE PROBLEM SITUATIONS TEST

Grade	Experimental	$Control_1$	$Control_2$	Total
Fourth	5.05	8.32	5.81	6.53
Fifth	4.95	3.32	4.50	4.19
Sixth (I)	4.74	7.00	6.63	6.14
Sixth (II)	5.95	4.82	4.56	5.12
Total	5.17	5.86	5.38	5.496

TABLE VI

ANALYSIS OF VARIANCE OF PRE-TEST SCORES ON THE PST

Source	d.f.	SS	MS	F	P
Treatments	2	20.86	10.430	0.580	> .20
Grades	3	188.89	62.963	3.516	< .02 >.01
T × G	6	205.78	34.297	1.915	< .10 >.05
Within Cells	216	3868.47	17.910	.....	
Total	227	4284.00	.....		

TABLE VII

MEAN PRE-TEST AND POST-TEST SCORES ON THE PST

Grade	Experimental		$Control_1$		$Control_2$		Total	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Fourth	5.05	3.74	8.32	6.64	5.81	7.47	6.53	5.89
Fifth	4.95	2.32	3.32	3.05	4.50	5.13	4.19	3.39
Sixth (I)	4.74	2.00	7.00	5.77	6.63	6.06	6.14	4.60
Sixth (II)	5.95	1.53	4.82	5.09	4.56	4.06	5.12	3.61
Total	5.17	2.39	5.86	5.14	5.38	5.67	5.496	4.373

group regressions is 29.639, the MS for deviations from group regressions is 5.527. The F-ratio is 5.363, which is significant below the .001 level for d.f.'s of 11 and 204. We thus reject the null hypothesis and conclude that heterogeneity of regression exists among the cells.

Regretfully, we are forced to abandon the analysis of covariance. Simply for purposes of completeness, it might be noted that the covariance analysis would not have changed any of the P-values in Table VIII very much.

Under the heading "Statistical analysis", two alternate procedures were suggested in the event that the covariance analysis proved inappropriate. The first of these was to interpret the analysis in Table VI, which shows no significant pre-test differences between treatment groups, indicating that the treatment groups were equated on the pre-test. Statistically, this is literally true since the arithmetic differences are random. With this interpretation, we may now consider the analysis of the post-test scores, as shown in Table VIII, as our critical test. This type of experimental analysis is quite common, in fact more common than covariance.

Table VIII shows that the differences among treatments and among grades are significant, the respective Fs being 23.4 and 7.625, both significant beyond the .001 level. The interaction is not significant. The next step is to test the differences among pairs of treatment means. These means will be found in Table VII. They are 2.39 for the experimental group, 5.14 for $Control_1$ and 5.67 for $Control_2$. The t-tests of differences between means are shown in Table IX.

The experimental group is clearly lower in mean score than either of the two controls. The control groups, however, do not differ from each other. This general finding is also true of the individual class means as shown in Table VII. In each case the experimental class has the lowest mean. In three of the four grade levels, $Control_1$ is lower than $Control_2$ though the differences are numerically small.

The second suggestion was a sign test in which each pair of pre-test and post-test scores are compared for direction of change. A plus would indicate that the post-test score was larger than the pre-test, a minus the reverse, and a zero, no change. A chi-square is then applied to the frequencies.

Table X shows the frequencies of plusses, minuses, and zeros for the sign test for the three treatment groups.

The chi-square obtained from Table X is 33.46, which is significant below the .0001 level for four degrees of freedom. The results are clearly in favor of the experimental group which has the largest number of minuses and the smallest number of plusses. $Control_1$ has the next largest number of minuses and the next smallest number of plusses. The trend is revealed more clearly by breaking down Table X into its three individual chi-squares, of which only two need be computed for our purposes. Comparing the experimental group with $Control_1$, we obtain a chi-square of 13.13, which is significant below the .005 level for d.f. = 2. Comparing $Control_1$ with $Control_2$, the chi-square is 8.21, d.f. = 2, and $P = <.02 >.01$. In other words, the experimental group appears to have been most affected by the treatment, $Control_1$ next most affected and $Control_2$ least affected. $Control_2$ in fact shows almost exactly the number of minuses that would be expected by chance alone.

Reliability of the PST—An estimate of test-retest reliability can be obtained by correlating the pre- and post-test scores for $Control_2$, the untreated control group. For this purpose we can utilize data from the abandoned covariance analysis, a procedure which will provide an overall r with systematic differences among grade means eliminated.

The test-retest correlation turns out to be .71. This is a respectable reliability with a group which includes a fair smattering of 9-year-olds. Furthermore, the hiatus between test and re-test was over six months and it is customarily no more than a week or two for test-retest reliabilities. The unusually long gap ordinarily has a tendency to attenuate the correlation.

The CT: Pre-Test—The mean pre-test scores for the CT are shown in Table XI.

The analysis of variance of the pre-test scores is shown in Table XII.

As in the case of the PST, the variance due to treatments is insignificant ($F = 1.734$) while that for grades is significant ($F = 11.146$, $P = <.001$). However, for the CT the interaction variance is also significant ($F = 2.752$, $P = <.03$), a disturbing occurrence, since it indicates that the sampling of classes is non-random. The source of the interaction seems obvious; the means for fourth and fifth grades in $Control_1$ and for the experimental sixth grade (I) are atypical when compared with means in the same level or in the same treatment.

Lack of randomness is not unexpected when intact school classes are assigned to treatments. It is an unfortunate happening in factorial design, but in this particular case we need not yet be overly concerned. If the treatment effects are exceedingly strong, it is quite possible that the adjusted post-test scores will not have a significant interaction. In that case all will be well. If, however, the interaction effect remains, then the within cells MS will no longer be the appropriate error term for testing the main effects and the analysis will be very coarse and probably unrevealing. [3]

The CT: Post-Test—The post-test means are shown in Table XIII. The pre-test means

TABLE VIII

ANALYSIS OF VARIANCE OF POST-TEST SCORES ON THE PST

Source	d.f.	SS	MS	F	P
Treatments	2	456.68	228.340	23.400	< .001
Grades	3	223.30	74.400	7.625	< .001
T × G	6	81.61	13.602	1.394	> .20
Within Cells	216	2107.82	9.758	.....	
Total	227	2869.31	.....		

TABLE IX

COMPARISONS OF TREATMENT GROUPS ON THE PST POST-TEST

Comparison	t	P
Experimental - $Control_1$	5.68	< .0001
Experimental - $Control_2$	6.20	< .0001
$Control_1$ - $Control_2$	1.03	.30

TABLE X

SIGN TEST ANALYSIS OF PRE- MINUS POST-TEST SCORES ON THE PST

	Plus	Minus	Zero	Total
Experimental	6	57	13	76
$Control_1$	27	48	13	88
$Control_2$	31	20	13	64

Chi-square = 33.46; d.f. = 4; P = < .0001

TABLE XI

MEAN PRE-TEST SCORES ON THE CAUSAL TEST

Grade	Experimental	$Control_1$	$Control_2$	Total
Fourth	12.53	15.76	14.50	14.40
Fifth	12.05	9.24	12.63	11.03
Sixth (I)	8.42	11.48	11.13	10.42
Sixth (II)	10.84	10.68	10.88	10.78
Total	10.96	11.79	12.28	11.658

TABLE XII

ANALYSIS OF VARIANCE OF PRE-TEST SCORES ON THE CT

Source	d.f.	SS	MS	F	P
Treatments	2	63.57	31.785	1.734	< .20 > .10
Grades	3	612.88	204.293	11.146	< .001
T × G	6	302.64	50.440	2.752	< .03
Within Cells	228	4178.89	18.328	.....	
Total	239	5157.98	.....		

TABLE XIII

MEAN PRE-TEST AND POST-TEST SCORES ON THE CT

Grade	Experimental		$Control_1$		$Control_2$		Total	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Fourth	12.53	5.53	15.76	12.44	14.50	12.56	14.40	10.28
Fifth	12.05	4.63	9.24	5.88	12.63	11.25	11.03	6.92
Sixth (I)	8.42	3.58	11.48	9.44	11.13	10.44	10.42	7.85
Sixth (II)	10.84	5.58	10.68	8.56	10.88	6.88	10.78	7.17
Total	10.96	4.83	11.79	9.08	12.28	10.28	11.658	8.054

are again included for comparative purposes.

All twelve of the classes show some decrement, all three treatment groups show reductions in mean score. $Control_2$ fell 2.00 points, $Control_1$ 2.71 points, while the experimental group dropped over six points, a decrease of more than 55 percent. The analysis of variance of post-test scores is shown in Table XIV.

The variance due to interaction is clearly still significant ($F = 5.031$, $P = < .001$). F-ratios and P-values for treatments and grades were computed using both the within cells MS and the T × G MS as error terms. The treatments MS is significant in either case, the respective Fs being 40.040 and 7.958, the respective Ps, $< .001$ and $< .025$.

The significance of differences among treatment groups is encouraging, but the persistent interaction is still a problem. There is not much point in testing for homogeneity of regression until we determine whether or not the interaction will remain significant when it is adjusted by covariance. Accordingly, the adjusted interaction MS and the adjusted within cells MS were computed. The results are shown in Table XV.

The interaction remains significant even after adjustment, the F-ratio being 7.956, $P = < .001$. This means that the within cells MS is no longer an appropriate error term for testing the main effects. The design would be left with only 10 degrees of freedom, 2 for treatments, 3 for grades, and 5 for interaction (since 1 d.f. is lost from the error term due to adjustment). Such an analysis could hardly be expected to provide significant results unless the treatments were practically infinitely powerful. One would hardly consider undertaking an experiment with only three scores in each treatment group.

Rather than forego the increased sensitivity of design offered by the within cells error, the data were inspected in the hope of discovering the source of the significant interaction. An examination of the data in Table XIII revealed that the sixth grade (II) class in $Control_2$ had dropped significantly on the post-test. Its pre-test mean was 10.88 and its post-test mean was 6.88. The t-score of the difference is 4.65, which is significant beyond the .01 level for 14 d.f. The difference of 4.00 points is more than twice that for any other class in $Control_2$ and greater than that for any class in $Control_1$, the treated control group. This class evidently contributes a considerable amount to the significance of the interaction. It does not seem conceivable that a single untreated control class should show a significant decrement. It is probable that this class had been exposed to some uncontrolled "treatment" during the course of the six months intervening between pre- and post-tests.[4] It was decided that sufficient grounds existed for dropping out this entire level from the analysis proper, if for no other reason than to determine statistically if this single class was, in fact, accountable for any large part of the interaction. The recomputed analysis of pre-test scores based on 9 classes and 180 subjects shown in Table XVI.

The results are almost identical with those of the original analysis of pre-test scores in Table XII. Treatment variance is insignificant and variances due to grades and interaction are significant. As in Table XII, if the T × G MS is used as the error term, the grades variance also becomes insignificant.

The analysis of post-test scores is shown in Table XVII.

Once again the results are practically unchanged. Except for minor discrepancies in P-values, Table XVII shows the same kind of data as did Table XIV, the original post-test analysis. All three effects are significant when tested by the within cells error; the treatment variance remains significant when T × G is the error term, while the grades variance becomes insignificant.

So far, the elimination of the sixth grade (II) level has not changed the analysis. The next step is to adjust the within cells MS and the T × G MS by covariance, as in the original analysis. The computations are shown in Table XVIII.

The F-ratio for the interaction is now only 0.994, which is clearly insignificant. A comparison of the results in Table XVIII with those of the original analysis in Table XV reveals the marked effect of the elimination of the sixth grade (II) $Control_2$ class. No other result is changed, but the interaction goes from highly significant to insignificant.

Homogeneity of regression must still be demonstrated before we can proceed to make the crucial adjustments of the treatment means. The MS for differences among group regressions is 14.093 and the MS for deviations from group regression is 8.577. The F-ratio for the test is $14.093/8.577 = 1.643$. The P-value is $< .20 > .10$ for 8 and 162 degrees of freedom. Hence, we may accept the null hypothesis and conclude that the cells have homogeneous regressions.

Having demonstrated homogeneity of regression, we may now proceed to adjust the sums of squares for treatments and grades for the crucial test. The adjusted data, plus the data of Table XVIII, are shown in Table XIX.

The adjusted variance for treatments yields an F-ratio of 53.933, which is significant beyond the .0001 level. The F for grades is 4.024, $P = .02$. There can be no doubt but that the treatment variance is significant in the final analysis and we must conclude that there have been real treatment effects during the six and one-half months intervening between pre- and post-testings. By comparing the unadjusted within cells MS (14.828) with the adjusted with-

TABLE XIV

ANALYSIS OF VARIANCE OF POST-TEST SCORES ON THE CT

Source	d.f.	SS	MS	F		P	
				Within	Int.	Within	Int.
Treatments	2	1213.22	606.610	40.040	7.958	< .001	<.025
Grades	3	425.55	141.850	9.363	1.861	< .001	>.20
T × G	6	457.37	76.228	5.031	.....	< .001	
Within Cells	228	3454.16	15.150	.....			
Total	239	5550.30	.....				

TABLE XV

COMPUTATION OF THE POST-TEST INTERACTION ON THE CT ADJUSTED BY COVARIANCE

Source	d.f.	SS	MS	F	P
T × G	6	440.21	73.368	7.956	< .001
Within Cells	227*	2093.39	9.222	.....	

*One degree of freedom lost due to adjustment

TABLE XVI

ANALYSIS OF VARIANCE OF PRE-TEST SCORES ON THE CT WITH GRADE LEVEL 6 (II) ELIMINATED

Source	d.f.	SS	MS	F	P
Treatments	2	85.47	42.735	2.395	.10
Grades	2	551.63	275.815	15.458	< .001
T × G	4	280.27	70.068	3.927	: 005
Within Cells	171	3051.18	17.843	.....	
Total	179	3968.55	.....		

are again included for comparative purposes.

All twelve of the classes show some decrement, all three treatment groups show reductions in mean score. $Control_2$ fell 2.00 points, $Control_1$ 2.71 points, while the experimental group dropped over six points, a decrease of more than 55 percent. The analysis of variance of post-test scores is shown in Table XIV.

The variance due to interaction is clearly still significant ($F = 5.031$, $P = < .001$). F-ratios and P-values for treatments and grades were computed using both the within cells MS and the T × G MS as error terms. The treatments MS is significant in either case, the respective Fs being 40.040 and 7.958, the respective Ps, $<$ .001 and $<$ .025.

The significance of differences among treatment groups is encouraging, but the persistent interaction is still a problem. There is not much point in testing for homogeneity of regression until we determine whether or not the interaction will remain significant when it is adjusted by covariance. Accordingly, the adjusted interaction MS and the adjusted within cells MS were computed. The results are shown in Table XV.

The interaction remains significant even after adjustment, the F-ratio being 7.956, $P = <.001$. This means that the within cells MS is no longer an appropriate error term for testing the main effects. The design would be left with only 10 degrees of freedom, 2 for treatments, 3 for grades, and 5 for interaction (since 1 d.f. is lost from the error term due to adjustment). Such an analysis could hardly be expected to provide significant results unless the treatments were practically infinitely powerful. One would hardly consider undertaking an experiment with only three scores in each treatment group.

Rather than forego the increased sensitivity of design offered by the within cells error, the data were inspected in the hope of discovering the source of the significant interaction. An examination of the data in Table XIII revealed that the sixth grade (II) class in $Control_2$ had dropped significantly on the post-test. Its pre-test mean was 10.88 and its post-test mean was 6.88. The t-score of the difference is 4.65, which is significant beyond the .01 level for 14 d.f. The difference of 4.00 points is more than twice that for any other class in $Control_2$ and greater than that for any class in $Control_1$, the treated control group. This class evidently contributes a considerable amount to the significance of the interaction. It does not seem conceivable that a single untreated control class should show a significant decrement. It is probable that this class had been exposed to some uncontrolled "treatment" during the course of the six months intervening between pre- and post-tests.[4] It was decided that sufficient grounds existed for dropping out this entire level from the analysis proper, if for no other reason than to determine statistically if this single class was, in fact, accountable for any large part of the interaction. The recomputed analysis of pre-test scores based on 9 classes and 180 subjects shown in Table XVI.

The results are almost identical with those of the original analysis of pre-test scores in Table XII. Treatment variance is insignificant and variances due to grades and interaction are significant. As in Table XII, if the T × G MS is used as the error term, the grades variance also becomes insignificant.

The analysis of post-test scores is shown in Table XVII.

Once again the results are practically unchanged. Except for minor discrepancies in P-values, Table XVII shows the same kind of data as did Table XIV, the original post-test analysis. All three effects are significant when tested by the within cells error; the treatment variance remains significant when T × G is the error term, while the grades variance becomes insignificant.

So far, the elimination of the sixth grade (II) level has not changed the analysis. The next step is to adjust the within cells MS and the T × G MS by covariance, as in the original analysis. The computations are shown in Table XVIII.

The F-ratio for the interaction is now only 0.994, which is clearly insignificant. A comparison of the results in Table XVIII with those of the original analysis in Table XV reveals the marked effect of the elimination of the sixth grade (II) $Control_2$ class. No other result is changed, but the interaction goes from highly significant to insignificant.

Homogeneity of regression must still be demonstrated before we can proceed to make the crucial adjustments of the treatment means. The MS for differences among group regressions is 14.093 and the MS for deviations from group regression is 8.577. The F-ratio for the test is $14.093/8.577 = 1.643$. The P-value is $< .20 >$ .10 for 8 and 162 degrees of freedom. Hence, we may accept the null hypothesis and conclude that the cells have homogeneous regressions.

Having demonstrated homogeneity of regression, we may now proceed to adjust the sums of squares for treatments and grades for the crucial test. The adjusted data, plus the data of Table XVIII, are shown in Table XIX.

The adjusted variance for treatments yields an F-ratio of 53.933, which is significant beyond the .0001 level. The F for grades is 4.024, $P = .02$. There can be no doubt but that the treatment variance is significant in the final analysis and we must conclude that there have been real treatment effects during the six and one-half months intervening between pre- and post-testings. By comparing the unadjusted within cells MS (14.828) with the adjusted with-

TABLE XIV

ANALYSIS OF VARIANCE OF POST-TEST SCORES ON THE CT

Source	d.f.	SS	MS	F		P	
				Within	Int.	Within	Int.
Treatments	2	1213.22	606.610	40.040	7.958	< .001	< .025
Grades	3	425.55	141.850	9.363	1.861	< .001	> .20
T × G	6	457.37	76.228	5.031	.....	< .001	
Within Cells	228	3454.16	15.150	.....			
Total	239	5550.30	.....				

TABLE XV

COMPUTATION OF THE POST-TEST INTERACTION ON THE CT ADJUSTED BY COVARIANCE

Source	d.f.	SS	MS	F	P
T × G	6	440.21	73.368	7.956	< .001
Within Cells	227*	2093.39	9.222	.....	

*One degree of freedom lost due to adjustment

TABLE XVI

ANALYSIS OF VARIANCE OF PRE-TEST SCORES ON THE CT WITH GRADE LEVEL 6 (II) ELIMINATED

Source	d.f.	SS	MS	F	P
Treatments	2	85.47	42.735	2.395	.10
Grades	2	551.63	275.815	15.458	< .001
T × G	4	280.27	70.068	3.927	: 005
Within Cells	171	3051.18	17.843	.....	
Total	179	3968.55	.....		

in cells MS (8.837) we see that the covariance analysis has nearly doubled the precision of the tests of the main effects, a result which makes the required time and effort well worthwhile.

We can now be certain that the treatment effects are significant, but we do not yet know which group or groups account for the significance. To investigate this point we first adjust the cell and treatment means and then apply t-tests to individual pairs. The adjustment of means is accomplished by the use of a regression equation which is derived from the covariance analysis.

Table XX lists the adjusted means for each class and for the various treatment groups. These means have had the effects of the respective pre-test means eliminated from them and will thus stand alone for comparison with each other without reference to the pre-test means.

The next step is compute t-tests for comparisons of pairs of treatment means. The three t-tests results are shown in Table XXI.

The t-tests reveal a clear-cut trend; the experimental group has the lowest adjusted mean score, differing significantly from both control groups; $Control_1$ has a significantly lower mean than $Control_2$. Returning to Table XX, we see that this trend holds true for all of the grade levels as well as for the treatment means.

The main analysis is now complete. It will be discussed in the next section.

Reliability of the CT—The test-retest reliability of the CT is .73. The remarks concerning the reliability of the PST also apply here.

Discussion and Conclusions—There is little doubt that the classes of the experimental teachers showed a marked change on both measures when compared with the classes of the control teachers. The statistical difficulties—the lack of homogeneity of regression for the PST and the peculiar interaction effect for the CT—do not obviate the large experimental-control differences.

It is perhaps unfortunate that the covariance analysis was inapplicable to the PST. There is, however, a plausible explanation for the heterogeneity of regression which precluded its use. Examination of Table VII shows that the post-test means for the experimental classes, especially the fifth grade and the sixth grades, are perilously near the ceiling (i.e., the lowest possible score) of the test. This means that considerable number of subjects obtained the same scores, mostly in the range 0-2. Since no one could improve beyond a score of zero, many of the subjects whose pre-test scores varied achieved a common post-test score. This tends to attenuate the pre-post correlation. This was, of course, not true for the control groups. The net result was that the experimental classes showed different regressions than the controls. The correlation between pre- and post-test scores was .74 for $Control_1$ subjects, .71 for $Control_2$ subjects, but only .44 for the experimental group.

Evidently the PST is an inadequate index for the pre-post type of experimental design since (a) the mean PST pre-scores are too low, and (b) the ceiling of the test is then too close to the pre-scores to permit adequate discrimination among members of experimental groups.

The performance of the $Control_1$ classes is not easily evaluated. This group of teachers was made acquainted with various teaching aids used by the experimental teachers and was permitted to use any that they wished in any fashion and for any amount of time. Only the briefest instructions concerning use of the materials were given since it was felt that such instructions fell within the province of the teacher training program. The purpose of the inclusion of $Control_1$ was to attempt to determine the effects of using the teaching materials uninstructed as opposed to the effects of the training program. The results are somewhat ambiguous. Two of the three analyses show that the $Control_1$ subjects improved significantly more than the wholely untreated $Control_2$ though not as much as the experimental subjects. Individual t-tests based on adjusted CT scores and individual chi-squares based on the sign test of PST results reveal this trend. The t-tests derived from the analysis of variance of PST results do not show this trend.

In an attempt to investigate this point further, the teachers in the experimental group and in $Control_1$ were asked to estimate the amount of class time spent in the use of the teaching aids. The amount of time in hours for each teacher is shown in Table XXII.

The experimental teachers used the teaching materials much more than did the control teachers, as would be expected. The experimental teachers also varied only slightly among themselves, as evidenced by the average deviation from the mean of 1.5 and the mean of 34 hours. On the other hand, the control teachers varied considerably, the average deviation being 4.5 and the mean 7 hours.

From the data in Table VIII we saw that there was no interaction between treatments and grades on the PST although both effects were significant. The same conclusions apply to the CT (Table XIX) when the sixth grade (II) level was eliminated. The lack of interaction means that the differences between experimental and $Control_1$ classes were about the same from grade level to grade level despite the significant overall differences in grade level scores. The difference between the experimental fourth grade class and the $Control_1$ fourth grade class is about the same as the difference between the two fifth grade classes, and so on.

TABLE XVII

ANALYSIS OF VARIANCE OF POST-TEST SCORES ON THE CT WITH GRADE LEVEL 6 (II) ELIMINATED

Source	d.f.	SS	MS	F		P	
				Within	Int.	Within	Int.
Treatments	2	1341.72	670.860	45.243	11.642	< .001	< .025
Grades	2	362.53	181.265	12.225	3.146	< .001	< .2 > .1
T × G	4	231.08	57.626	3.886	.....	.005	.....
Within Cells	171	2535.62	14.828	.....			
Total	179	4470.95	.....				

TABLE XVIII

COMPUTATION OF THE POST-TEST INTERACTION ON THE CT WITH GRADE LEVEL 6 (II) ELIMINATED, ADJUSTED BY COVARIANCE

Source	d.f.	SS	MS	F	P
T × G	4	35.13	8.738	0.994	> .20
Within Cells	170*	1502.25	8.837	.....	

*One degree of freedom lost due to adjustment

TABLE XIX

ANALYSIS OF COVARIANCE OF THE POST-TEST CT SCORES WITH GRADE LEVEL 6 (II) ELIMINATED

Source	d.f.	SS	MS	F	P
Treatments	2	953.21	476.605	53.933	< .0001
Grades	2	71.12	35.560	4.024	.02
T × G	4	35.13	8.738	0.994	> .20
Within Cells	170*	1502.25	8.837		

*One degree of freedom lost due to adjustment

TABLE XX

POST-TEST MEAN SCORES ON THE CT, ADJUSTED BY COVARIANCE

Grade	Experimental	$Control_1$	$Control_2$
Fourth	5.192	10.223	11.076
Fifth	4.572	7.457	10.854
Sixth	5.634	9.714	10.917
Total	5.132	9.131	10.951

TABLE XXI

COMPARISONS OF TREATMENT GROUPS ON THE CT ADJUSTED POST-TEST SCORES

Comparison	t	P
Experimental - $Control_1$	7.603	< .0001
Experimental - $Control_2$	9.793	< .0001
$Control_1$ - $Control_2$	3.315	.001

TABLE XXII

CLASSROOM HOURS SPENT USING TEACHING AIDS

Group	Fourth	Fifth	Sixth (I)	Sixth (II)	Mean	Average Deviation
Experimental	33	33	37	33	34	1.5
$Control_1$	14	2	9	3	7	4.5

The data in Table XXII show that the experimental classes were all treated approximately alike with respect to number of hours of use of teaching aids. The control classes, however, varied from 2 hours to 14 hours. If the use of teaching aids alone had any real effects, we would expect to have found a variation in differences between pairs of classes. The two fourth grades, for example, should not differ as much as the two fifth grade classes since the fourth grade Control$_1$ teacher spent seven times as much time with teaching aids as did the fifth grade Control$_1$ teacher. This variation would have resulted in a significant interaction between treatment and grades. Since no such interactions were found with the PST or with the CT after elimination of the sixth grade (II) level[5], we conclude that amount of class time spent using the teaching materials was probably not the sole factor making for reduction in test scores, even if we accept the conclusion that there was a significant change in the scores of the subjects in Control$_1$ classes. That we should accept this conclusion is still open to question, as is the matter of what factor did influence the scores of the subjects in Control$_1$ if we do not accept it. We have gone as far as we can reasonably go with the present analysis. Data are not available to settle either question satisfactorily.

Summary—The subjects of this investigation were four classroom teachers and their pupils, each classroom matched with two control groups. The teachers participated in a training program designed to extend their understanding and appreciation of child behavior, to provide opportunity for growth in personal adjustment and to develop methods for teaching causally oriented curricular content.

Tests of the child's awareness of the complex multiple causative nature of human behavior and of his tendency to immediate punitiveness were administered to both experimental and control groups in the fall of the school year and again approximately 6-1/2 months later.

Extended analyses of the results using both parametric and non-parametric methods, as the nature of the data indicated, were applied.

The classes of the experimental teachers showed distinctly significant changes on the two measures used when compared with classes of the control teachers.

It thus appears that when we bring children of the upper elementary grade levels under the influence of causally oriented teachers teaching causal content we bring about significant differences in the child's growth in the aspects measured in this study.

Additional differences between causally oriented and control subjects will be reported in other papers.

REFERENCES

1. Dixon, W. H. and Massey, F. J. Introduction to Statistical Analysis (New York: McGraw-Hill Co., 1951).
2. Levitt, E. E. "Punitiveness and 'Causality' in Elementary School Children," Journal of Educational Psychology (in press).
3. Levitt, E. E. and Lyle, W. H. "Evidence for the Validity of the Children's Form of the Picture-Frustration Study," Journal of Consulting Psychology, 1955 (in press).
4. Levitt, E. E. and Ojemann, R. H. "The Aims of Preventive Psychiatry and 'Causality' as a Personality Pattern," Journal of Psychology, XXXVI (1953), pp. 393-400.
5. Lindquist, E. F. Design and Analysis of Experiments (New York: Houghton Mifflin Co., 1953).
6. Lyle, W. H. and Levitt, E. E. "Punitiveness, Authoritarianism and Parental Discipline of Grade School Children," Journal of Abnormal and Social Psychology, 1955 (in press).
7. Ojemann, R. H. "An Integrated Plan for Education in Human Relations and Mental Health," Journal of National Association of Deans of Women, XVI (1953), pp. 101-108.
8. Stiles, Frances S. A Study of Materials and Programs for Developing an Understanding of Behavior at the Elementary School Level, Ph.D. Dissertation, University of Iowa, 1947.
9. Stiles, Frances S. "Developing an Understanding of Human Behavior at the Elementary School Level," Journal of Educational Research, XLIII (1950), pp. 516-524.
10. Zelen, S. L. Effect of a Causal Learning Program, Mimeographed Report, Preventive Psychiatry Project (Iowa City: State University of Iowa, 1954).

FOOTNOTES

1. All of the analyses of variance computed for this report are based on Lindquist's procedures (5). "Treatments" refers to the three primary groups, the experimental and the two controls. "T × G" is the treatment-by-grades interaction. The "within cells" mean square is the overall standard error when other systematic differences have been eliminated. If the interaction is not significant, the mean square for within cells is the appropriate error term for testing the effects. In the remaining tables of this kind, "sum of

squares'' will appear as SS and ''mean square'' as MS.

2. The assumptions necessary for the use of covariance will be found in (5).

3. Strictly speaking the within cells MS is not the appropriate error term in Table XII, although it was used there to test the main effects. If the T × G MS, which is really the appropriate error term, had been used, the respective F-ratios for treatments and grades would be 0.636 and 4.05, the latter falling just short of the .05 level of significance. Since the treatments MS simply remains insignificant and since we are not interested in grade differences, it is immaterial which error term is used.

4. A lengthy interview with the teacher by the experimenter most familiar with her (Whiteside) did not provide any clues as to what this uncontrolled factor might have been.

5. The interaction did not actually involve the sixth grade (II) class in $Control_1$ but was rather a function of this class in $Control_2$. Comparing only the experimental and $Control_1$ groups, there was no significant interaction even with the sixth grade (II) class included.

THE SELECTION OF CANDIDATES FOR TEACHER EDUCATION AT THE UNIVERSITY OF WISCONSIN

GUSTAVE JOHN STOELTING *
Milwaukee Public Schools

SECTION I

BACKGROUND OF THE PRESENT INVESTIGATION

A. Basic Principles in Screening of Candidates for Teaching

MANY NEW screening procedures have come to be used by teacher training institutions during the last decade in an effort to select more students capable of becoming superior teachers. This seems important if our schools are to have competent leadership. Today the majority of teacher-education schools employ some form of screening of persons admitted or educated as teachers.

This strong interest in teacher selection arises out of a three-fold need:

1. To maintain high standards in the profession at a time when emergency measures to overcome teacher shortages may permit standards to fall.
2. To find more individuals equal to the increasingly complex task of teaching.
3. To prevent the wastes created when individuals are trained for positions for which they are personally or intellectually not qualified.

The teacher shortages of the past twelve years have given rise to emergency measures that permit large numbers of poorly qualified individuals to become teachers. While teacher training institutions and professional organizations have attempted to overcome shortages with programs of recruitment, merely encouraging larger numbers to become teachers is not an adequate solution to the problem. To have good teachers it is necessary to exercise a certain amount of selection from among the larger groups of interested individuals who choose to become teachers. The result has been a greater variety of screening devices and their more widespread use.

The literature on teacher selection repeatedly stresses the importance of protecting the public welfare through careful selection of candidates for teacher education. New concepts of learning and development of young individuals combine to make classroom management a highly intricate procedure. The necessity of helping young people understand a complex environment and the demands that it makes on the individual emphasize further the necessity for more competent teachers. Individuals with specific qualities are frequently called to meet the requirements of special situations. Stiles (30) summarizes this point as follows:

> Superficial consideration might lead one to believe that democratic principles would compel institutions to admit all who desire to become candidates for teacher education. More objective thought, however, would help one to realize that since education is a function of the state and is maintained for its own good, therefore, the state has not only the right but also the responsibility to secure the best possible teachers. Merely providing state institutions of higher learning with competent professors and adequate curricula will not assure the state that superior teachers will be developed. The type of teacher that the university or teachers college will ultimately produce is dependent upon the quality of persons who are accepted for training.

Much of what has been done in construction of devices for screening of teacher candidates has been based on investigations of factors involved in successful teaching. The factors most commonly used in screening teacher candidates are intelligence and scholastic achievement. La Duke (19), Rostker (26), and Seagoe (27) independently investigated the relationship between intelligence and success in teaching. They found a significant, positive relationship. As measures of intelligence have become generally more reliable, the use of this device for screening of teacher candidates has become almost universal.

Lins (20) and Stuit (31) provide data on the relationship between scholastic achievement and success in teaching. Most teacher training institutions today specify a minimum of scholastic achievement as a part of their program of

*The author wishes to express his appreciation to Dr. C. S. Liddle, Dr. G. G. Eye, and Dr. A. S. Barr for helpful criticisms and suggestions in the planning and carrying out of the study.

screening for teacher candidates.

In addition to using intelligence and scholastic achievement as measures for predicting future teaching success, some teacher training institutions are using also less well known measures of yet other factors in teaching success. Flanagan (11) and Seagoe (27) provide data supporting the use of a general culture test for screening prospective teachers. The importance of proficiency in reading and speech as vital qualities of the successful teacher is supported by data provided by Flanagan (11), Henrickson (14), and McCoard (21).

Personality as a factor in teacher success has stimulated much interest and lively discussion. The use of personality measures in teacher selection is on the increase. More general use of such measures is limited because both satisfactory rating scales and standards of teacher personality are lacking. Some institutions now using this factor in a screening device do so only to discover instability in a candidate. Experimental screenings of candidates using personality devices are under study in a number of teacher training institutions.

B. General Features of Screening for Teacher Selection

The screening devices described in this section and their placement in a program of selection represents a composite of practices as reported in the literature on teacher selection. There is much variation among teacher training institutions in this respect.

Admission is the crucial point in most teacher selection programs. This is true largely because failure to predict the success of a candidate at the time of admission may lead to great waste. As a result, a large number of devices are in use by teacher training institutions to screen for admission. Screening devices may be divided into two catagories: application and orientation.

Screening through application generally is based on information on the educational and home background. The most frequently used data are the high school record, scholastic attainment, personality and attitude ratings of the applicant by school staff members, standardized test scores, and participation in extra-curricular activities. To complete the data on educational background the administrative head of the school from which the applicant is graduated is ordinarily requested to make some sort of a statement regarding the applicant's general acceptability for continued training.

Sometimes the information on educational background is supplemented by a wide variety of information on home and personal background. Such questions as age, occupation, and educational attainment of various members of the family are asked. The applicant is also sometimes requested to furnish an autobiography, a report of financial responsibility, a statement of purposes for continuing his education, and personal references regarding his acceptability.

Admission is usually followed by a period of intensive testing to aid in orientation. Some institutions prefer to have the results of such testing in hand before the candidate's application for admission is acted upon. While the latter arrangement has some obvious disadvantages, it does provide much additional data to assist in making the vital decision made at the time of admission.

Testing programs at the time of admission generally include such areas as English placement, general culture, intelligence, interests, personality, and reading. Within these areas there is wide variety of instruments used.

The greatest variation in the selection and use of test instruments lies in the area of personality where there appears to be little agreement on the qualities of a good teacher. Among the instruments frequently used are the Minnesota Multiphasic Personality Inventory, the Bell Adjustment Inventory, and the Bernreuter Personality Inventory. The use of a subjective technique, the group interview, and projective techniques are being explored by several teacher training institutions.

Two devices other than standardized tests are also commonly used in screening for teacher selection; the physical examination ordinarily seeks to determine not only whether an individual is capable of undertaking a normal course of studies, but also if he has any physical defects which might limit his efficiency as a teacher and thus increase the element of risk involved in accepting him as a candidate.

Speech tests have a dual purpose. They are used to eliminate those applicants whose speech defects are such as to be a definite handicap in the profession. They also disclose remediable defects in applicants otherwise qualified.

Applications for admission generally are reviewed by an admissions official responsible for weighing the quality of each applicant as a student and prospective teacher in the light of all the evidence available. As screening procedures have become more refined a few teacher training institutions have turned the important task of evaluating an applicant's qualifications over to a committee. Such a move emphasizes the importance attached to the evaluation of an individual's potential success as a teacher when application is made for admission.

A second important point in the teacher selection programs of many teacher training institutions occurs at the end of the second year of preparation, or the beginning of the third year. Here it takes the form of "Admission to Profes-

sional Study" or "Admission to Senior College", or it may simply constitute a more intensive stage in evaluation of the candidate's qualifications in a continuous screening process. By and large, no matter what the name or what procedures are used, at this point the candidate's progress is examined as to his suitability as a teacher.

The most important new data commonly used at this point is the record of a candidate's academic achievement and the pattern of credits earned. Academic achievement as measured by the Grade Point Average or its equivalent is used by many schools to maintain minimum standards. Of all devices for screening reported in the literature, the maintenance of minimum scholastic standards appears most frequently.

To insure uniform training deemed essential for successful teachers many teacher training institutions specify an academic pattern to be followed by their candidates. While this device does not provide specifically for screening, it again sets basic requirements which the institution deems necessary for teacher success.

Some teacher training institutions employ personal interviews, a physical examination, and a speech test, at this point, to determine admission to teacher training program rather than at the time of admission to the school. A few institutions use these devices both at the time of admission and after the first two years of training.

As further evidence of a candidate's progress in becoming a teacher some schools examine the type of activities in which the candidate has engaged beyond the requirements of the curriculum. Emphasis is placed upon participation in those activities which appear to contribute to teacher success following graduation. The use of this device, as in the case of specific curricular requirements, does not necessarily constitute screening in that it selects the better candidate and eliminates the poorer. It does, however, set requirements which must be met, thus exposing the candidate to further experiences and training on which he may draw later as a well-qualified teacher.

The pattern of evaluation already discussed is often combined with evaluation of achievement in professional training courses and practice teaching. Thus a continuous process of evaluation covering the entire period of professional training is provided.

In general, it would appear that much of the screening practices of the final two years of training are different from those that are commonly employed at the time of original admission. Most teacher training schools do the large share of screening at the time of admission; the promising candidates are admitted to school and the poor risks are eliminated. Thereafter it would appear that the function of screening changes. If a candidate continues to make normal progress in meeting the increasing requirements of training and skill, the screening process "selects" him for further training. On the other hand, a candidate is eliminated who does not have sufficient interest to fulfill the screening requirements satisfactorily or if a defect serious enough to impair successful teaching is discovered.

The final stage of screening by teacher training schools is graduation, certification, and placement. Most schools of education combine graduation and certification, i.e., the school certifies the individual as a teacher when he graduates. Some few schools draw a finer distinction between graduation and certification. These schools will graduate a candidate, provided he has fulfilled all the requirements; he may have achieved only the minimum professional standards required by law, but because he has not attained the standards set by the school, the candidate is not certified. To gain certification at such schools candidates must attain higher qualifications. The institution, in turn, certifies to the professional competence of the individual.

Placement is not generally regarded as a part of the screening program. However, teacher training institutions are sensitive to its screening function. Screening controls are restricted or relaxed as opportunities for placement change. In addition, the use of screening devices indicates the desire by schools for teacher education to reduce waste by eliminating those who might have difficulty in being placed.

C. Survey of the Literature

The literature on teacher selection may be divided into four areas:

1. General Philosophy

Teacher selection is based on the important principle that good schools depend on well qualified teachers. Such standards of competency in turn depend upon good training, a background of accepted research, and capable individuals. But good training can be had only as research findings in the field are put in practice. On this basis the literature recognizes that the better selection of candidates is the key to higher professional standards.

It is further pointed out in the literature that better teachers are needed for the increasingly complex job of teaching. As the center of learning moves away from the highly organized fields of subject matter toward the needs of the individuals, the task of guiding and directing the learning process places increasing demands upon the teacher. In order to discharge such an important task adequately, more capable individ-

uals are needed.

The literature also points to the necessity for reducing the human and social waste involved in training and employing individuals who are not well suited to the task of teaching. The development and use of good selection procedures is a means of avoiding such waste, while at the same time providing for more efficient use of the teacher-training facilities.

Comprehensive statements of basic philosophy regarding teacher selection are found in articles by Flowers (12); Kirkpatrick (16), and Morris and Phillipson (23). The latter article is particularly valuable in that it describes the areas of research necessary to develop more adequate teacher-selection procedures.

2. General Surveys of Teacher Selection Procedures

Studies on the prevailing practices in teacher selection in limited areas of the United States have been reported by Haskew (13), and Stiles (30). Their conclusions may be summarized as follows:

a. Most institutions have a teacher selection program.
b. The programs range from simple requirements to highly developed procedures in the process of being further refined.
c. The element of timing in selection procedures varies from a single, "on the spot" selection made only once to a continuous selection process beginning while still in high school and extending to graduation and placement.
d. Selection is the responsibility of one person in most institutions. Some few have a committee.
e. Most common bases of selection are:
   Scholastic achievement
   High school record
   Results of aptitude tests
   Results of interviews
f. Personality traits are being studied and used increasingly in teacher selection programs.
g. Some teacher training institutions without a selection program feel that as public institutions they do not have the right to exclude.
h. Most of the institutions without a selection program do not have one because of an apparent lack of reliable bases to make a selection.

3. Descriptions of Existing Programs of Selection

Reports of specific programs of teacher selection are common in the literature. Four of these reports bear mention here for they represent the most advanced practices in teacher selection. The selection program in the Connecticut Teacher's Colleges is described by Engleman and Larson (10); the plan in use in New Jersey is reviewed by West (32); the San Diego State College selection program is described in an article by Alcorn (1); and the development and operation of the teacher selection program at Syracuse University in New York is given by Smith (28), and White (33). The significant feature of each of these teacher selection plans is that they utilize subjective techniques of personality analysis in addition to other more common sources of information to aid in making a judgment.

4. Summary of Review of Literature

Comprehensive reviews of the literature on teacher selection have been written by Archer (3), Barr (4), and Haskew (13). From these articles the following conclusions may be drawn:

a. There is a lack of reliable objective data on which teacher selection may be based.
b. The reviews report several studies to show significant correlations between success in teaching and scholarship.
c. Increasing use is being made of tests of aptitude for various special fields.
d. Speech tests are becoming more common in programs of teacher selection.
e. Greater emphasis is being placed on personality in teacher selection. Most of the teacher training institutions using this factor employ it to detect the unstable. Some few institutions report the use of experimental techniques to select candidates who demonstrate desirable personality traits in a social situation.
f. Evidence of leadership qualities are becoming increasingly important as a part of teacher-selection programs.
g. Tests of proficiency in basic skills are being used to supplement intelligence tests.
h. Committee selection procedures are steadily replacing selection by a single individual.
i. There is an increasing recognition that teacher selection cannot depend on a single factor, but must be based on a constellation of factors.

SECTION II

STATEMENT OF THE PROBLEM

A. Selection and Teacher Success

THE CENTRAL problem of this study is to determine the efficiency with which the several selective devices employed at the University of Wisconsin operate in choosing potentially successful teachers out of the total group seeking admission and eventual certification for teach-

ing. To do this the study will seek to answer five questions:

1. How well do present selection procedures discriminate between the superior teacher candidate and the teacher candidate who is likely to meet with only limited success?
2. Under what circumstances do selection devices now employed permit admission of individuals not likely to succeed?
3. Is there basis for raising or lowering the standards by which candidates are admitted to preservice training and certification as teachers?
4. At what point in the teacher education program is the screening for teacher education likely to be most effective?
5. What recommendations, based on the findings of the study, can be made for improved procedures for the selection of candidates for teaching?

To study the effectiveness of the screening devices used at the University of Wisconsin, the data on which selection of the 1952 graduating class of the School of Education was based will be related to success of the individuals of the class following graduation. These data include:

a. Rank in high school class
b. Psychological scores
   Henmon-Nelson
   American Council on Education
c. Cooperative Reading test score
d. Cooperative General Culture test score
e. Predicted Grade Point average
f. Earned Grade Point average
g. Minnesota Multiphasic Personality Inventory test score
h. Speech proficiency test score

These data will be correlated to the various criteria for measuring success in teaching. The criteria will consist of:

a. An in-service rating by the principal or superintendent of those who were employed in a teaching situation during the year since graduation.
b. A departmental rating based on the estimate of a candidate's effectiveness as a teacher by the faculty of his major department.
c. A Placement Bureau rating based on the candidate's general acceptability as a teacher.
d. Practice-teaching grades.

These ratings will first be considered separately after which they will be combined into a single rating for each individual included in the study.

A measure of the efficiency of the selection procedures used by the School of Education will be obtained through correlating the scores used in screening with the criteria of teaching success. By this means it will be possible to determine how well the screening devices can discriminate between teachers of superior, average, and inferior teaching ability. The information gained through this study should offer a basis for improving the screening procedures in the School of Education of the University of Wisconsin, and also provide a means for continuous evaluation of the program.

B. Selection of Candidates for Teacher Training at the University of Wisconsin

The selection of candidates for teacher training at the University of Wisconsin has a dual purpose: (1) to assure that all individuals who are accepted for training as teachers will succeed, and (2) to assure that a larger proportion of those accepted for training are capable of becoming superior teachers. Thus screening seeks to protect individuals from entering a field of work in which they may not succeed, while at the same time protecting our schools by supplying better teachers.

A major point in the screening of candidates for teacher training at the University of Wisconsin occurs at the time of admission. The data on which admission is based includes personal data (physical characteristics, appearance, interests, ambitions), family data (nationality, parent's occupation, residence, siblings), educational background (academic record, test record, pattern of credits earned, personality rating, extra-curricular activities) and a statement by the administrator of the preparatory school regarding the educational promise of the individual.

The data is evaluated by an official in the Admissions Office at the University of Wisconsin. Greatest emphasis in the evaluation is placed upon the future academic promise of the applicant. Upon admission each student is assigned to the school of his choice and an advisor in his major field. A student who expresses a preference for entering the School of Education is enrolled as "Pre Ed", and assigned to an advisor in the College of Letters and Science.

During the week of registration new students participate in a program of orientation to life at the University. An important feature of this program is the extensive testing done during the period. The tests included in the program are the Cooperative Reading Test, the Cooperative General Culture Test, and the American Council on Education Psychological Examination. The results derived from these tests are used to advise and counsel the student during his first two years at the University. These test results also have an important function in the screening of teacher candidates at the time of admission

to professional study.

Following admission to the University, there is no direct screening of teacher candidates until the student applies for transfer to the School of Education at the end of the fourth semester of study. Two basic requirements must be met during the first four semesters to be admitted to professional study in the School of Education: (1) a student must have earned at least 62 credits of an approved course of study with a minimum 1.3 grade point average; and (2) the course of study a student presents for evaluation at the end of four semesters' work must meet the standard requirements for majors and minors, specific course requirements, and requirements varying according to the major and minor departments.

At the end of the fourth semester of study (or when 62 credits in an approved pattern have been earned) the student may apply for transfer to the School of Education for professional training. Evaluation of a student's record up to that point constitutes a second major point in the screening process. Data on which the screening is based includes a transcript of credits earned, grade point average, high school rank, and the results from the orientation tests taken during the registration period at the beginning of the first semester at the University.

The most important factors in the screening are the two basic requirements for admission to the School of Education—completion of course requirements and maintenance of a 1.3 grade point average.

Course requirements which must be completed before an applicant may be admitted to professional study include:

a. English attainment requirements
b. Physical Education or Military Science
c. Minimum requirements in majors and minors
d. A minimum of 62 credits

In addition each major department has varying requirements which the individual must meet.

In some cases when most requirements have been met and the candidate presents records otherwise suitable, he may be admitted on the condition that certain deficiencies will be removed during the following semester. In other cases where many requirements remain to be completed, the candidate must utilize an additional semester or summer session before application for transfer may be made.

The record of credits earned is also evaluated for grade point average (based on 1 grade point per credit for a final course grade of "C", 2 grade points per credit for a final course grade of "B", and 3 grade points per credit for a final grade of "A"). A minimum total grade point average of 1.3 is specified for admission to the School of Education.

Candidates whose application for admission to teacher training is rejected on the basis of a grade point average too low to meet the minimum requirement may request to have his case reviewed by the Dean of the School of Education, or an assistant. In such a case, compensating factors such as an above-average I.Q., or above average high school rank, are sought in the candidate's records. Such candidates whose records are otherwise satisfactory may be admitted on a strict probationary basis.

While data from the Cooperative General Culture and Cooperative Reading tests are used as a part of the screening process, it does not play a part in a candidate's admission to the School of Education. These data are used to aid the individual candidate and his advisor in plotting the most appropriate course of professional study based on his skills and interests.

Following admission to professional study the student remains subject to course requirements while maintaining the 1.3 grade point average. Both of these devices continue to serve the screening function in that they eliminate those who cannot reach the minimum standards of success in teacher training.

During training the candidates must meet three other screening situations to qualify for graduation and certification as a teacher. The first of these is a speech test which is administered jointly by the School of Education and the Department of Speech. Its purpose is to certify that the speech proficiency of the teacher candidate is of a satisfactory standard for classroom work. Provision is made for remedial work for those who cannot qualify on the initial test. Occasionally this device may screen out such individuals whose speech handicaps are such as to limit their efficiency in the classroom.

The Minnesota Multiphasic Personality Inventory serves as a second screening device during the period of professional training of teachers. Use of the inventory is limited to the detection of such individuals whose personality is unstable to the point of limiting their effectiveness in the classroom. Such individuals are referred to the Student Health Clinic for treatment and are counseled into other fields of work.

Finally, a candidate must present a certificate of physical health and fitness from the University Medical Examiner as an indication that no physical defects exist to limit the individual's success as a teacher.

When the candidate has successfully met each of these screenings the School of Education is willing to certify his success as a teacher by granting the University Teacher's Certificate. Through the use of the screening devices as described only those whose success as teachers is reasonably assured are retained for training and

graduated with certification.

SECTION III

GATHERING THE DATA

A. The Study Group

AS A BASIS for study of the selection procedures employed by the School of Education of the University of Wisconsin, the 1952 graduating classes (February, June and August) were chosen for study. Members of these classes have been teaching one or more years, thus giving an adequate basis for in-service success rating.

The combined membership of these three classes is 352; 134 were men and 218 were women. A preliminary survey of the group made in October, 1953, disclosed that 54 men and 133 women, or a total of 187, taught during the first year following graduation; a total of 165 did not teach, —80 were men and 85 women. Of the 80 men who did not teach, 30 were in military service; of the 85 women who did not teach 27 were married.

The remaining non-teaching graduates may be grouped as follows:

1. Attended graduate school—19 men, 9 women
2. Decided not to teach—12 women
3. No record of employment, and no reply to two inquiries—13 men, 9 women
4. Entered private industry—11 men, 12 women
5. Other public employment—4 men, 4 women
6. Unplaced—3 men, 9 women

While this non-teaching group appears large, it is possible that many may eventually become teachers. Some of those now in service and others in the Graduate School will doubtless enter the profession later. Nevertheless considering the totals involved, the non-teaching group appears large.

Further study of the 1952 graduating group made it apparent that much of the data used for screening was not available for those who transferred to the University of Wisconsin after a year or two of study elsewhere. These were not processed by the usual admission procedures, nor were the data of the orientation testing program available. Therefore, only those who had originally entered the University as freshmen and who had gone through the entire procedure of admission and screening were included in the study. Thus, 163 transfers to the University of Wisconsin were dropped from the study for lack of data, leaving 189 in the group to be studied.

The placement records for this group provide the data presented in Tables I, II, and III.

B. Methods Employed in Gathering the Data

To facilitate gathering of the data, a special 4" × 6" card was devised and printed for use in the study. One card was prepared for each individual. On the top line of the card beginning with the left margin the name of the individual was typed. The space immediately below the name was reserved for the date of entry into the University and a notation whether the individual was an original entry or a transfer student. The upper right hand corner was used to record the date of graduation and the individual's academic major and minors. The space below Criterion Rank was used to record the details of the individual's placement.

Since all test data were filed according to date of entry, the transcript of each student's record was the logical starting point. Transcripts of the graduates were made available by the School of Education Dean's office. In addition to the date of entry the transcript also contained high school rank and earned grade point average data. Since rank in high school class had already been converted to a percentile score, these data were simply transferred to the record card. To compute the earned grade point average it was necessary to count the number of credits and grade points earned and to record them in fraction form to be calculated later.

With the date of entry available, the gathering of test data could go ahead since this data was filed according to the student's entry date. The test data included:

1. Henmon-Nelson Psychological
2. American Council on Education Psychological
3. Cooperative Reading
4. Cooperative General Culture

The information on these tests was made available through the Student Counseling Center. Since the data for each of the tests were already in percentile rank form, the data were transferred directly to the individual record for each graduate in the study group. The Student Counseling Center also furnished data on each individual's predicted grade point average (based on a regression equation using high school rank and percentile rank from the American Council on Education Psychological examination to predict Grade Point Average).

Inasmuch as only the raw scores were available for the Minnesota Multiphasic Personality Inventory, it was necessary to complete a profile and code for each member of the study group before the individual's score could be recorded.

TABLE I

SUMMARY OF PLACEMENT OF THE 1952 STUDY GROUP OF THE SCHOOL OF EDUCATION AT THE UNIVERSITY OF WISCONSIN
(Survey of October, 1952)

Men employed in teaching positions	21	
Women employed in teaching positions	77	
		98
Men employed in non-teaching positions	42	
Women employed in non-teaching positions	49	
		91
Group Total		189

TABLE II

SUMMARY OF PLACEMENT OF THE 1952 STUDY GROUP OF THE SCHOOL OF EDUCATION AT THE UNIVERSITY OF WISCONSIN IN TEACHING POSITIONS
(Survey of October, 1952)

Teaching Field	Men	Women	Total
Agriculture	2		2
Art Education	2	5	7
Business Education		1	1
Chemistry	1	1	2
Economics		1	1
English	1	10	11
French		2	2
Geography	1		1
History	3	1	4
Home Economics		26	26
Mathematics		1	1
Music		9	9
Natural Science	2	1	3
Physical Education	7	5	12
Recreation		7	7
Sociology	1		1
Speech		1	1
Speech Correction	1	6	7
Total Men Teaching	21		
Total Women Teaching		77	
Total Graduates Teaching			98

TABLE III

SUMMARY OF PLACEMENT OF THE 1952 STUDY GROUP OF THE SCHOOL OF EDUCATION AT THE UNIVERSITY OF WISCONSIN IN POSITIONS OTHER THAN TEACHING
(Survey of October, 1952)

	Men	Women	Total
Decided Not to Teach	1	6	7
Graduate School (U of W)	5	2	
(U of Chicago)	1		8
Married		17	17
Military Service	21	1	22
No Reply	7	7	14
Other Public Employment	1	4	5
Private Industry	6	5	11
Unplaced		7	7
Total Men Not Teaching	42		
Total Women Not Teaching		49	
Total Graduates Not Teaching			91

TABLE IV

CORRELATIONS OF FOUR CRITERIA OF TEACHING SUCCESS WITH TEST DATA EMPLOYED IN SCREENING CANDIDATES FOR TEACHER TRAINING

	Criteria of Teaching Success			
Screening Data	In-Service Rating	Departmental Rating	Placement Bureau Rating	Practice Teaching Grades
Henmon-Nelson Psychological	.314	.216	.119	.094
ACE Psychological	-.027	.163	.073	.026
Reading	.056	.240	.106	.059
General Coop. Culture				
Social Problems	-.169	.105	-.032	-.060
History	-.245	.101	.054	-.038
Literature	-.549	.087	-.112	-.039
Science	-.176	.045	.000	-.061
Fine Arts	-.042	-.161	-.133	-.194

The data on the speech screening test were made available in the office of Professor Gladys Borchers, Chairman of the Education-Speech Committee. Ratings in their original form were: A - superior, B - above average, and C - average (no student is certified with less than a "C" rating, and is assigned remedial work until a "C" rating is earned). To give these ratings numerical basis for statistical purposes, an "A" was recorded as "5", "B" as "4", and "C" as "3".

With the exception of the data from the Minnesota Multiphasic Personality Inventory, all the data was in a form readily adaptable to use in a correlational study. The MMPI data were not amenable to such a study.

During the time screening data was being gathered, the data on criteria of teaching success to which screening data will be related was also being recorded. The criteria of teaching success include:

1. An in-service rating
2. A departmental rating
3. A Placement Bureau rating
4. Practice teaching grades

To obtain the in-service rating for the 98 graduates of the study group who were employed as teachers during the first year following graduation, a postal reply rating card was devised and printed (see Appendix F).* The rating is based on the individual's performance in his first year in teaching.

These cards were mailed to the superintendents or principals of 95 teachers in the group. Ratings for three of the teaching group who were employed as teachers of recreation by the American Red Cross were not requested because of much shifting of assignments, and no current information on what their present situation was; furthermore these individuals were not assigned in one location long enough to give an accurate in-service rating. Wherever possible, the request for a rating was sent directly to the superintendent or principal in charge.

Within 14 days of the mailing date, 76 (80%) had been returned. At the end of 30 days, 88 (93%) had been returned. Two of the remaining seven for whom no rating was returned had not been placed as the survey of placement had indicated. The remaining five were placed out of the State of Wisconsin, and, lacking the name of their principal or superintendent, no further effort was made to obtain an in-service rating.

The in-service ratings obtained through this means were recorded as "5" for a superior rating, "4" for an above-average rating, "3" for an average rating, "2" for a below-average rating, and "1" for an inferior rating. By giving these ratings a numerical value it was possible to make various statistical analyses of them.

A second criterion of teaching success consists of a departmental rating. This rating is made by the faculty of each individual's major department. The department's rating is an estimate of the individual's potentialities as a teacher. Since a major department's most important contact with the individual is through his classwork, the rating may reflect heavily the individual's academic achievement in his major subjects.

The departmental ratings for the 1952 graduating classes were not uniform in the type of ratings employed. To produce as much uniformity between the departmental ratings as possible each department prepared a key for translating the scores into superior, above average, average, below average, inferior ratings. These, as with the in-service ratings, were recorded as numerical quantities ("5" for a superior rating, "4" for an above average rating, etc.,) for direct use in the computations.

A Placement Bureau rating was the third criterion of teaching success. This rating, made by the Assistant Director of the Placement Bureau, depends on a group of factors not likely to appear in the other criteria ratings. The following factors were said to be involved in arriving at a rating:

1. Credentials—statements of observing officials, advisors, teachers, and supervisors regarding the individual's promise. Other information used here includes statements by the candidate himself regarding his interests, preferences, and ambitions.
2. Observations—appearance, attitudes and general adjustment of the individual is observed in a personal conference, in connection with his routine duties, and in social situations.
3. Reviews of practice teaching performance by the critic teachers.
4. Transcript is consulted for placement purposes only; it is not used for rating purposes. Grade point average is used for rating purposes only when very high or very low.
5. The departmental rating is considered only when very high or very low.

Ratings were provided on the superior, above average, average, below average, inferior scale. The ratings were recorded on the same numerical basis as the other ratings.

Only 141 ratings could be provided by the Placement Bureau since 48 in the study group did not register with the Bureau. No special at-

* All references to Appendices may be found in original thesis filed in the Library, University of Wisconsin, Madison, Wisconsin.

tempt was made to determine why these 48 did not register, but a simple survey of the records disclosed that a large proportion were those who decided not to teach, those women who were married and decided not to seek placement, and the graduates who majored in Recreation and were placed through other placement facilities.

A final criterion of teaching success consists of practice teaching grades. These grades are based on each individual's attainment in two practice teaching situations—one semester of practice teaching in a minor academic field and one semester of practice teaching in the major academic field. These grades do not appear separately on the transcript but are available separately in the Student Teaching records office. With separate major and minor practice teaching grades available for each individual, the grades were averaged to produce a single practice teaching grade.

Practice teaching grades are generally awarded on a superior, average, inferior basis using an A, B, C grading system. It was necessary, therefore, to assign the numerical values to these grades as follows:

A (superior)	5
A-, B+ (above average)	4
B (average)	3
B-, C+ (below average)	2
C or below (inferior)	1

With these numerical values the ratings will be used in the computations in the same way as the other criteria.

A single, over-all criterion of teaching success was derived from an average of the four criteria described above. No weighting was given the separate criteria: (1) since one individual was responsible for each of the ratings, it is felt that the judgment of any individual should not be emphasized more than the others; (2) with a straight average to produce the criterion, no one factor in teaching success is emphasized. It is felt that all the factors involved in arriving at the criteria ratings are contributory to teaching success, and should be considered equally.

In computing the criterion, 80 of the total study group had all four of the criteria available. An additional 66 of the total group had three criteria available to formulate their criterion. For the remaining 43 of the total group only two criteria operated in arriving at their criterion of teaching success.

To avoid a marginal criterion of teaching success ratings for seven individuals, it was necessary to give emphasis to a single rating. Wherever these occur emphasis was given in the direction of the in-service rating, if available; to the Placement Bureau rating if the in-service rating was missing; or to the practice teaching grades if both the in-service ratings and the Placement Bureau ratings were not available.

These criteria of teaching success were considered separately in correlation studies with the screening data, and then, combined into a criterion of teaching success, were correlated again with the screening data.

SECTION IV

ANALYSIS OF THE DATA

A. The Criteria of Teaching Success

TO GET further data relative to the criteria of teaching success, intercorrelations were calculated among them.

The correlation between the in-service ratings and the departmental ratings, based on 88 cases for whom in-service ratings were available, was .319.

It is entirely probable that these ratings have only academic ability as a common element. The department faculties were decidedly limited in the aspects of teaching upon which their estimates could be based. Academic ability was the one aspect with which these individuals were most familiar. The in-service ratings depended upon this and other qualities as well.

While there is little relationship between these data it is felt that both areas covered by the ratings are of importance in the training of a teacher as well as in success in teaching.

The highest correlation between any two criteria of teaching success was .627 for 84 cases based on the in-service and Placement Bureau ratings. A strong similarity in what the ratings attempt to measure doubtless accounts for the relatively high correlation. In both ratings the academic record is consulted, but not emphasized. Furthermore, in both the ratings, personality becomes a matter of considerable importance. Such matters as adjustment in social situations, interest in people, attitudes toward community responsibilities, and general personal appearance are important factors in both the in-service and the Placement Bureau ratings.

The relationship between the in-service ratings and practice teaching grades, based on 88 cases, was .327, which is low. It is interesting that there should be so little in common in the two ratings. It is possible that a teacher's ability to organize and discharge a set of duties comprising an actual teaching position is different from that provided by practice teaching. It would seem on the basis of the low correlation here derived that a study should be made of the factors producing such a result.

The correlation involving 141 cases between the departmental ratings and Placement Bureau

ratings was .472.

While both ratings make use of academic achievement as part of the rating, the Placement Bureau would appear to have recognized the importance of the individual's personality in teaching.

Further differences in the ratings emphasize the supplementary character of the ratings. An individual's ability to adapt himself to a specific job, to a school organization, and to a community is of much importance in the Placement Bureau rating, while the departmental rating is not so much concerned with this factor. A third important difference in the ratings concerns the type of performance each is concerned with; the Placement Bureau rating is alert to performance in leadership and organization of social services while the departmental rating depends largely on academic performance.

A correlation of .551 involving 189 cases indicates considerable similarity between the departmental rating and practice teaching grades. It appears likely that both of these ratings depend heavily on academic achievement.

It appears that the relatively high correlation of departmental-practice teaching ratings may offer some clue to the inability of practice teaching grades to predict in-service success. Greater similarities occur between practice teaching grades and departmental ratings than practice teaching grades and in-service ratings. It is likely, then, that greater emphasis in practice teaching grades is being placed on the academic aspects of teacher preparation rather than leadership and organizational factors considered necessary to succeed on the job.

The Placement Bureau rating serves as a transitional rating between training for teaching and actual teaching in the field. This rating correlates well with measures of in-service success, emphasizing the practical aspects of teaching, and also correlates well with measures of teacher success taken during preparation for teaching, emphasizing the theoretical and academic aspects of teaching. A measure of teaching success involving the factors used in the Placement Bureau rating warrants further investigation for possible adaptation to pre-training selection purposes.

The value of the departmental ratings and practice teaching grades seems to be low for purposes of predicting in-service success. These ratings probably emphasize the academic and theoretical factors in teacher training as a basis for measuring teacher success. As a result of the emphasis, however, they reflect teacher success in training, and justify their use as criteria of success.

Since all of the elements used to arrive at the criteria ratings are of importance in some phase of teacher preparation and teaching, and teacher selection needs be concerned with all these aspects, a straight average of the criteria is used to produce a composite criterion of teacher success. It is assumed that this criterion of teacher success will reflect all these important elements in weighing the ability of a screening device to discriminate between levels of teaching ability.

In an effort to get further data on the interrelationships of the criteria of teacher success a multiple R was calculated using the departmental rating, Placement Bureau rating, and practice teaching grades to predict in-service success. The R was .629. Comparison of this figure with the intercorrelation figures on the teaching success criteria will show that the three ratings used together to predict in-service success are no better than the rating used by the Placement Bureau alone. It further indicates that what is being measured in the departmental ratings and practice teaching grades has no particularly significant relationship to in-service teaching success.

B. Correlations of Screening Data With the Criteria

The data used for screening may conveniently be divided into three groups; standardized test data, academic achievement data, and speech proficiency test data. Table IV shows the correlations of the standardized test data with the four criteria of teaching success. Table V shows the correlations of the same test data with a criterion of teaching success.

Examination of the correlations on Table IV and V, the relationship between standardized test data and the criteria of teaching success, discloses that only two correlations in both tables are very different from zero. The first of these, namely, the correlation between the Henmon-Nelson psychological scores and in-service success, is quite low. The other correlation, the -.549 between literature scores and in-service success ratings, would not generally be acceptable as evidence of teacher acceptability.

It must not be assumed that the evidence given in Tables IV and V is proof that the areas covered by the tests are not important in successful teaching. Even the least well prepared may appear adequate to rating officials, —thus little or no correlation. Then, too, these particular instruments possibly cannot be relied upon to screen teacher candidates. Other instruments in the same areas may be able to perform the screening function adequately where the instruments under study here have failed.

Reference to Table VI will show that data on academic achievement is promising for use in screening. While the correlations are not high, the data can be used with a reasonable de-

TABLE V

CORRELATIONS OF A CRITERION OF TEACHING SUCCESS WITH TEST DATA EMPLOYED IN SCREENING CANDIDATES FOR TEACHER TRAINING

Screening Data	Criterion of Teaching Success
Henmon-Nelson Psychological	.139
ACE Psychological	.103
Reading	.172
General Cooperative Culture	
Social Problems	-.054
History	-.005
Literature	-.035
Science	.003
Fine Arts	-.206

TABLE VI

CORRELATIONS OF FOUR CRITERIA OF TEACHING SUCCESS WITH ACADEMIC ACHIEVEMENT DATA EMPLOYED IN SCREENING CANDIDATES FOR TEACHER TRAINING

Screening Data	In-Service Rating	Departmental Rating	Placement Bureau Rating	Practice Teaching Grades
High School Rank	.221	.205	.199	.237
Predicted Grade Point Average*	.047	.309	.166	.115
Earned Grade Point Average (4 semesters)*	.385	.335	.302	.375

*Correlation between Predicted and Earned Grade Point Average = .570.

gree of validity.

It will be noted that the correlations of High School Rank with the various criteria of teaching success are low; when High School Rank is correlated with the composite criterion of teaching success, the correlation becomes worthy of consideration. It is doubtful, however, if the r .270 is high enough to justify the use of High School Rank as a screening instrument. Possibly its use may be justified if the severe limitations imposed by the low validity are observed.

Predicted Grade Point Average does not appear to qualify for use as a screening device. Only the correlation with the departmental rating is considerable. Its correlation with Earned Grade Point Average is .570, which is always a consideration in the training of teachers.

The Earned Grade Point Average has correlations with the criteria which are somewhat higher, more consistent and probably more useful as a screening instrument.

Correlation with the criterion is higher than with each of the criteria separately, indicating that the Earned Grade Point Average can predict moderately well over a wide range of measures of teaching success.

Since Earned Grade Point Average is the one screening device capable of discriminating between levels of teaching ability, its use might possibly be broadened to include other devices. Earned Grade Point Average is now used as a basis for admission to professional study in the School of Education, a 1.3 grade point average being the minimum. This might well be carried on through the final two years of preparation. A separate requirement might be set up to apply to professional courses. A minimum required 1.8 grade point average, for example, could be used to screen teacher candidates for higher professional standards, while a 1.3 grade point minimum could remain for all other courses. Such adaptations could broaden the use of the only valid screening included in this study.

Reference to Table VIII above will show that the Speech Proficiency test is not capable of predicting teacher success, there being a near zero relationship between speech scores and success in teaching. This certainly does not mean, however, that the speech test no longer serves an important function. The low correlations probably arises out of the fact that extreme cases have been removed from the teacher preparation program or that the deficiency has been overcome. Those that meet this standard seem adequate.

The use of the speech proficiency test as a screening device probably should be continued to insure minimum speech proficiency. As such, it will not be necessary to rate the individual, but merely to certify that he meets minimum standards, or to withhold certification until he becomes qualified through remedial work.

In order to further describe the efficiency with which the various selective devices operate the correlations between the selective devices and the various criteria were converted into an efficiency score through the use of the Predictive Efficiency formula. In this formula, Pre. Eff. $= 1 - \sqrt{1-r^2}$ where $\sqrt{1-r^2}$ is the Coefficient of Alienation. The coefficient gives a basis for decinding how high a correlation must be in order to be satisfactory for predictive purposes(24: 115). This subtracted from 1 gives a decimal fraction which can be treated as a percentage. A predictive efficiency percentage above 90 is regarded as high, between 10 and 90 as moderate, between 5 and 10 as low, and below 5 as negligible.

The only correlations whose predictive efficiency was better than 5% were between the earned grade point average and the criteria of teaching success. The correlation between earned grade point average and the criterion of teaching success yielded a 9% predictive efficiency; between earned grade point average and the in-service rating, 8% predictive efficiency; between earned grade point average and the departmental rating, 6% predictive efficiency; and between earned grade point average and practice teaching grades, 7% predictive efficiency. All other predictive efficiency scores were less than 5%, thus not reliable for predictive purposes.

C. The Minnesota Multiphasic Personality Inventory and Teaching Success

The Minnesota Multiphasic Personality Inventory was included as a part of the screening program by the School of Education in order to detect individuals with personalities such as to limit their effectiveness in the classroom. The data used in this study concerns only those whose code score met the standards considered to be adequate for teaching. Accordingly, with their elimination those remaining should be adequate as shown by subsequent results.

The data, when classified according to the categories of teaching success (namely, Superior, Above Average, Average, Below Average, and Inferior, based on the criterion of teaching success), yielded no discernable personality patterns. Personality codes which appeared among teachers judged to be inferior were found in equal or greater proportion among the other categories of teaching success. Furthermore, personality codes indicating a mild maladjustment appeared as frequently among the average, above average, and superior teachers as was the case among the below average or inferior.

Further investigation in the use of this instrument is possible, but beyond the scope of the present investigation. The responses on the

TABLE VII

CORRELATIONS OF A CRITERION OF TEACHING SUCCESS WITH ACADEMIC ACHIEVEMENT DATA EMPLOYED IN SCREENING CANDIDATES FOR TEACHER TRAINING

Screening Data	Criterion of Teaching Success
High School Rank	.270
Predicted Grade Point Average	.207
Earned Grade Point Average (4 Semesters)	.407

TABLE VIII

CORRELATIONS OF A SPEECH PROFICIENCY TEST USED IN SCREENING CANDIDATES FOR TEACHER TRAINING WITH FOUR CRITERIA AND A CRITERION OF TEACHING SUCCESS

Screening Data	Criterion of Teaching Success
Speech	.179
Speech—In-service Rating	.011
Speech—Departmental Rating	.253
Speech—Placement Bureau Rating	.069
Speech—Practice Teaching Grades	.221
Speech—Composite Criterion of Teaching Success	.179

test may be analyzed item by item to determine what items, if any, are able to discriminate between different levels of teacher success. Thus, while the test as a whole is not a valid screening instrument, separate items within the test may be found entirely valid for use in screening. The data collected on the candidates for teacher training through the Minnesota Multiphasic Personality Inventory indicates that this test is incapable of predicting teacher success.

D. Conclusions of the Study

On the basis of the data described in the foregoing pages, answers are proposed to the basic questions involved in this study:

1. How well do present selection procedures discriminate between the superior teacher and the teacher likely to meet with only limited success?

On the basis of present selection procedures none of the standardized tests used appear capable of predicting future teacher success. These include the Henmon-Nelson psychological test, the American Council on Education psychological test, the Cooperative Reading test, and the Cooperative General Culture test. Since the relation between scores earned on these tests, and eventual success in the profession are as low as they are, these tests would appear to eliminate both potentially successful teachers as well as unsuccessful.

Academic achievement data holds some promise for screening of teacher candidates, and the standards might be increased in this respect, but this will need to be done with care. Earned grade point average appears to be the most useful instrument in this group, and in the entire screening program for that matter, for predicting teacher success. As has been suggested earlier, the use of the overall grade point average may be broadened to include other devices for screening, in addition to raising or lowering the minimum as the occasion demands.

The use of High School Rank for screening purposes as far as the data here presented appears of doubtful value. Although the correlation of this device is larger than most, it appears low for predictive purposes particularly after a preliminary selection has been made on the basis of grade point average. Its use should probably be restricted to that of providing supplementary data.

The use of the Speech Proficiency Test should be continued in the screening program, at least for certification. It is important that teacher candidates be certified for minimum speech attainment necessary for classroom success. It is probably not necessary to rate candidates above the minimum requirements.

2. Under what circumstances do selection devices now employed permit admission of individuals not likely to succeed?

Of the 189 total group studied, only 24 were judged on the basis of the criterion to have achieved less than average success. This group was obviously permitted to enter training and become teachers in spite of the screening procedures employed. Since no follow-up, other than the in-service rating, was conducted on the group it is not possible to determine why these 24 met with limited success.

Of the total below-average group seven were admitted to professional study with earned grade point averages well below 1.3. An additional 12 were admitted with grade point averages between 1.30 and 1.50. These two sub-groups constitute 79% of the total below-average group, indicating that an academic basis exists for their low rating.

However, 4 in the below-average group had earned grade point averages above 2.00. Thus, it appears that while the earned grade point average is a valid measure of teaching success, it is not sufficient in and of itself. Further, experimental study will be necessary to discover other valid factors in teachirg success to be combined with Earned Grade Point Average in an improved screening program, capable of isolating those individuals not likely to succeed in teaching.

3. Is there basis for raising or lowering the standards by which candidates are admitted to pre-service training and certification as teachers?

If the minimum grade point average were increased from the 1.30 now being employed to 1.50, the higher minimum would screen out 13 of the 24 who were judged to be of less than average teaching ability. But at the same time, such an increase would eliminate 31 who were rated as average, 7 rated above average, and 1 rated superior. Thus, it becomes evident that change of the grade point average minimum will not alone be the solution to more adequate screening.

4. At what point in the teacher education program is the screening for teacher education likely to be most effective?

It is apparent that prediction of teacher success becomes easier and more accurate as more

information about the candidate becomes available. The most accurate prediction can be made at the time of graduation and certification, based on a 22% predictive efficiency of the Placement Bureau ratings. But since it is important for the efficient use of time and facilities to make a prediction of success as early as possible, a balance must be effected. Thus, the ideal time occurs when the decision to admit or reject is made early enough to allow a rejectee ample time to choose a new course without a great loss of credits, and late enough to determine the earned grade point average on which a reasonable judgment on future success in teaching may be based.

There is no decisive evidence on which selection of a point of most effective screening may be based. It is possible that the time of application for admission to professional study may be most effective, subject to further study.

It must be pointed out here that no adequate screening program will function properly with only one on-the-spot screening. Such screening must be supplemented by continuous selection procedures both before and after admission to professional study. These would include active supervision of academic progress and the course of study, periodic counseling, a speech proficiency test, a personality test, a physical examination, interview, and standardized tests. A program such as this would be effective because it allows time to gather adequate information about an individual on which to base admission, and at the same time providing for increasing standards of attainment necessary for well-qualified teachers.

5. What recommendations, based on the findings of the study, can be made for improved procedures for the selection of candidates for teaching?

a. It is entirely possible that standardized tests are now available which might be used in the screening program to replace those not now competent for use in screening. The literature on screening of candidates for teacher training gives evidence of many standardized devices now in use, though there is no conclusive proof of their relation to teacher success.

b. The use of the Grade Point Average may be broadened to include a specific minimum for professional courses. Since Grade Point Average was demonstrated as an effective screening device, standards of more intensive preparation may be possible with a device such as this.

c. The use of subjective techniques might be adapted to use for screening purposes. Techniques such as the group interview, group dynamics situations, and observation under social pressure are now in use and under study by a number of teacher training institutions. While their use seems promising, continuous research to check on the results is necessary before they can be depended upon.

d. As an aid in the study of experience and personal characteristics in a teacher candidate, a record system following the individual through his four years of preparation for teaching may be useful. In teacher training institutions now using this device on an experimental basis they find that observations going back into secondary school make significant contributions in the screening of successful teacher candidates.

e. Further research on the characteristics of the successful teacher are needed on how personal cultural pattern, philosophy, and system of values combine in a successful teacher.

BIBLIOGRAPHY

1. Alcorn, M. D. "The Problem of Teacher Selection," Educational Administration and Supervision, XXXIV (March 1948), pp. 160-62.
2. Almy, H. C. and Sorenson H. "A Teacher Rating Scale of Determined Reliability and Validity," Educational Administration and Supervision, XVI (March 1930), pp. 179-86.
3. Archer, C. P. "Personnel Procedures in Teacher Training Institutions," Journal of Educational Research, XL (May 1947), pp. 672-84.
4. Barr, A. S. "The Measurement and Prediction of Teaching Efficiency: A Summary of Investigations," Journal of Experimental Education, XVI (June 1948), pp. 205-83.
5. Barr, A. S. and others. "The Validity of Certain Instruments Employed in the Measurement of Teaching Ability," in The Measurement of Teaching Efficiency (New York: Macmillan Co., 1935), pp. 71-141.
6. Bliss, W. B. "How Much Ability Does a Teacher Need?" Journal of Educational Research, VI (June 1922), pp. 33-41.
7. Boardman, C. W. Professional Tests as Measures of Teaching Efficiency in High School, Contributions to Education, No. 327 (New York: Teachers College, Columbia University, 1928).
8. Breckenridge, E. "A Study of the Relation of Preparatory School Records and Intelligence Test Scores to Teaching Success," Educational Administration and Supervision, XVII (November 1931), pp. 649-60.
9. Broom, M. E. "The Predictive Value of Three Specified Factors for Success in Practice Teaching," Educational Administration and Supervision, XV (September

1929), pp. 25-29.
10. Engleman, F. E. and Larson, V. M. "Selective Admission to the Teaching Profession," NEA Journal, XXXIX (February 1950), pp. 94-95.
11. Flanagan, J. C. "An Analysis of the Results from the First Annual Edition of the National Teachers Examinations," School and Society, LIV (July 26, 1941), pp. 59-64.
12. Flowers, J. G. "Better Teachers for Our Schools," Peabody Journal of Education, XXV (January 1948), pp. 139-141.
13. Haskew, L. D. "Selection, Guidance and Preservice Preparation of Students for Public School Teaching," Review of Educational Research, XXII (June 1952), pp. 175-181.
14. Henrikson, E. H. "Comparisons of Ratings of Voice and Teaching Ability," Journal of Educational Psychology, LIV (February 1943), pp. 121-123.
15. Jacobs, C. L. The Relation of a Teacher's Education to Her Effectiveness, Contributions to Education, No. 277 (New York: Teachers College, Columbia University, 1922).
16. Kirkpatrick, F. H. Helping Students Find Employment, American Council on Education Studies, Series VI, Student Personnel Work No. 12 (Washington, D.C.: American Council on Education, 1949).
17. Knight, F. B. Qualities Related to Success in Teaching, Contributions to Education, No. 120 (New York: Teachers College, Columbia University, 1922).
18. Kriner, H. L. "Five-Year Study of Teacher's College Admissions," Educational Administration and Supervision, XXIII (March 1937), pp. 192-199.
19. LaDuke, C. V. "The Measurement of Teaching Ability," Journal of Experimental Education, XIV (September 1945), pp. 75-100.
20. Lins, L. J. "The Prediction of Teaching Efficiency," Journal of Experimental Education, XV (September 1946), pp. 2-60.
21. McCoard, W. B. "Speech Factors as Related to Teaching Efficiency," Speech Monograph, XI (January 1944), pp. 55-64.
22. Norris, E. H. Personal Traits and Success in Teaching, Contributions to Education, No. 342 (New York: Teachers College, Columbia University, 1929).
23. Norris, B. S. and Phillipson, H. "The Development of Research: Selection and Training of Teachers," Times Educational Supplement, 1630 (July 27, 1946), p. 351.
24. Peters, C. C., and Van Voorhis, W. R. Statistical Procedures and Their Mathematical Bases (New York: McGraw-Hill Book Co., 1940).
25. Retan, G. A. "Emotional Instability and Teaching Success," Journal of Educational Research, XXXV (October 1943), pp. 135-141.
26. Rostker, L. E. "The Measurement of Teaching Ability," Journal of Experimental Education, XIV (September 1945), pp. 52-74.
27. Seagoe, M. V. "Standardized Tests in the Pre-Training Selection of Teachers," Journal of Educational Research, XXXVI (May 1943), pp. 678-693.
28. Smith, H. P. "The Selection of Students for the Profession of Teaching," School and Society, LXV (March 8, 1947), pp. 169-171.
29. Somers, G. T. "Pedagogical Prognosis," Predicting the Success of Prospective Teachers, Contributions to Education, No. 140 (New York: Teacher College, Columbia University, 1923).
30. Stiles, L. J. "Recruitment and Selection of Prospective High School Teachers by Universities," Educational Administration and Supervision, XXXII (February 1946), pp. 117-121.
31. Stuit, D. B. "Scholarship as a Factor in Teaching Success," School and Society, XLVI (September 1937), pp. 382-384.
32. West, R. L. "The Operation of a Selective Admissions Program in a Teachers College," Educational Record, XXX (April 1949), pp. 137-147.
33. White, V. "Selection of Prospective Teachers at Syracuse University," Journal of Teacher Education, I (March 1950), pp. 24-51.

DIFFERENTIAL METHODS OF SOLVING SELECTED PROBLEMS ON THE ACE PSYCHOLOGICAL EXAMINATION*

LEONE ANDERSON, RICHARD RANKIN, JOY RICHARDSON,
JULIUS SASSENRATH, JULIUS THOMAS
University of California at Berkeley

TWO EXPERIMENTERS using eye movement records have attempted to uncover differential problem solving methods employed by high and low performers. Anselmo (2), using Number Series problems, and Greening (4), employing Figure Analogies problems, found that high performers took less time than low performers in the solution of the problems. Both experimenters concluded that the average duration of fixation pauses was slightly less for high performers, but the difference in methods of attacking the problems remained undetected. In an effort to improve and expand these earlier investigations, an attempt is made to analyze more thoroughly the eye movement data, and also to obtain verbal recordings of subjects during their solution of the problems.

Problem

The experimenters seek to disclose the differential problem solving processes employed by high and low performers, respectively, in the solution of Number Series and Figure Analogies problems from the ACE Psychological Examination (1). The following specific questions were investigated:

a. Do high performers exhibit fewer fixations and regressions than low performers?

b. Do high performers have a total duration of fixations and regressions, respectively, which is less than the low performers?

c. Do high performers fixate more on the pattern of the problem than on the options; i.e., do they establish the pattern before selecting an option as an answer, whereas do low performers have an equal number of fixations on the pattern and options of the problem?

d. Do high performers on Number Series problems shift more readily than low performers from one arithmetic process required to solve a problem to another process for a subsequent problem; i.e., shifting from addition to a combination of addition, multiplication, and division?

Methodology

The ACE Psychological Examination, in accordance with the instructions (1) was administered to two classes in Introductory Educational Psychology. From this population of 220 students those who missed less than six or more than 19 of the 30 problems on the ACE Number Series Sub-test were defined as the high and low performers, respectively, on Number Series. Those students who missed less than eight or more than 19 of the 30 problems on the ACE Figure Analogies Sub-test were defined as the high and low performers, respectively, on Figure Analogies. This method of selecting extreme performers ultimately provided the following number of subjects in the four groups: low Number Series (N = 5); high Number Series (N = 9); low Figure Analogies (N = 6); and high Figure Analogies (N = 11).

Eye movements were photographed by a corneal reflection type camera. The developed film was projected on a replication of the problems and the data were then tabulated. A detailed description of the camera and procedure is given by Gilbert (3).

A disc audograph** was used to record the subjects' verbalizations of the step by step progression of what they perceived and thought in their attempt to solve the problems. Data were compiled from the verbal recordings by developing a rating scale with dichotomous ratings (+ or -).

The following 11 items comprised the rating scale used in evaluating verbal responses to the Figure Analogies problems:

1. Identifies similarities only
2. Identifies differences only
3. Identifies similarities and differences
4. Uses mathematical concepts
5. Proceeds from an incomplete and/or inaccurate recognition of constants (those aspects of the figure which remain the same) and variables (those aspects of the figure which change)

* This experiment was conducted under the guidance of Professor Luther C. Gilbert, in the Educational Psychology Laboratory at the University of California, Berkeley.

**Gray Audograph, Hartford, Connecticut.

6. Develops an answer only through examination of the rule; then proceeds to options (Solution by Analysis)
7. Develops an answer by elimination of options (Solution by Elimination)
8. Does not apply relationship once having described it
9. Answers with correct solution, and (a) corrects error, (b) does not correct error, (c) makes no error
10. Answers with incorrect solution, although (a) corrects error, (b) does not correct error
11. Presents no solution

The following 13 items comprised the rating scale used in evaluating verbal responses to the Number Series problems:

1. Identifies numbers only
2. Identifies arithmetic relationship only
3. Identifies numbers and arithmetic relationships
4. Proceeds by insightful technique; i. e., notes rules after verbalizing only one to four numbers and/or relationships; then answers
5. Proceeds by repetitious technique; i.e., (a) notes five to seven numbers and/or relationships; then answers, (b) re-examines the numbers and/or relationships already verbalized; then answers
6. Develops an answer only through examination of the pattern; then proceeds to options (Solution by Analysis)
7. Develops an answer by elimination of options (Solution by Elimination)
8. On first examination proceeds from an incomplete and/or inaccurate recognition of the rule
9. On first examination notes new arithmetic relationships not present in preceding problems
10. Requires more than one examination to note new arithmetic relationships not present in preceding problems
11. Answers with correct solution and (a) corrects error, (b) does not correct error, (c) makes no error
12. Answers with incorrect solution, although (a) corrects error, (b) does not correct error
13. Presents no solution

The problems to be used with the eye movement camera and verbal recorder were chosen from different editions of the ACE in order to minimize the effect of practice. The nine problems to be solved before the camera were selected by an item-difficulty analysis of the Number Series and Figure Analogies Sub-tests of the 1941 edition of the ACE. Assuming that the same problems in the 1940 edition of the ACE had the respective degree of difficulty, problems were selected from this edition for use with the verbal recorder. For purposes of analysis the individual problems were divided into two parts: (a) the pattern, i.e., the initial part of the problem establishing the rule, and (b) the options, i.e., the multiple choice selection of answers. The problems, shown at top of next page, were chosen and classified as easy, medium, or difficult.

The laboratory procedure was introduced to each subject with a brief explanation of the purpose of the experiment and the principles of the camera and audograph. Each examinee was then (a) adjusted before the camera, (b) given instructions adapted from the ACE, (c) asked to solve two practice problems, and (d) individually presented the nine problems as an untimed test. Upon completing the camera procedure the subject was (a) seated before the audograph, (b) given instructions adapted from the ACE, (c) presented two practice problems to be solved verbally, and (d) administered the problems as an untimed test, to which the subject responded by verbalizing thought processes in attempting a solution.

Discussion of Results

From Table I it can be seen that the mean number of fixations and regressions for Number Series problems is less for high performers than low performers. This appears to differ with Anselmo's (1) findings, but may be accounted for in the more stringent selection of subjects and problems for this experiment. Mean total duration of fixations and regressions for Number Series problems also indicates that the high performers spend less time than low performers. However, the mean number of correct answers for Number Series problems is not very different for the low and high performers. Thus without controlling the time variable, the two groups performed equally well. This was not true when these subjects were administered the ACE as a timed test from which they were selected as high and low performers, respectively.

Different results emerge (Table I) for the two groups tested on Figure Analogies problems. Except for mean number of correct answers, there appears to be little or no difference between any of the measures. This lack of differences between high and low performers is contrary to Greening's (4) conclusions. The apparent difference between mean number of correct answers indicates that even on an untimed basis the high performers are superior. Thus the Figure Analogies problems tended to function as a power test, while this was not true of the Number Series problems. The design of the Number Series and Figure Analogies problems does not, of

	Easy	Medium	Difficult
(Camera) Number Series (ACE 1941)	2, 4, 5	16, 17, 18	28, 29, 30
(Camera) Figure Analogies (ACE 1941)	2, 4, 9	16, 20, 21	28, 29, 30
(Audograph) Number Series (ACE 1940)	2, 4, 5	16, 17, 18	28, 29, 30
(Audograph) Figure Analogies (ACE 1940)	2, 4, 9	16, 20, 21	28, 29, 30

course, permit a comparison of the data on these two types of problems.

Figure 1 indicates that for the three levels of difficulty of the Number Series and Figure Analogies problems, the high performers fixated more on the pattern than did the low performers. Moreover, both groups fixated more on the pattern than on the options of all the problems with the exception of the low performers on Figure Analogies.

As the Number Series and Figure Analogies problems increased in level of difficulty, from easy to medium, both the high and low performers spent a similarly greater percentage of fixations establishing the pattern of the problem. Specifically, $\triangle FA_1$, approximates $\triangle FA_2$, and $\triangle NS_1$ approximates $\triangle NS_2$ (Figure 1). However, with those problems which increased from a medium to a difficult level, the low performers fixated less on the pattern, thus relying more on an eliminative method selecting an answer from the options. In contrast, the high performers in their solution of the difficult problems increased the number of their fixations on the pattern, thus indicating a more analytic method of problem solving. Here the differential between $\triangle NS_2$ and $\triangle NS_3$ is less than that between $\triangle FA_2$ and $\triangle FA_3$.

Presumably, the effectiveness of the high performers' method of problem solving is not appreciably reduced by the increasing difficulty levels of those problems administered. Rather it demonstrates a proportionately greater intensity of analysis. (Intensity is here taken to be a function of the percentage of fixations on the pattern.)

Low performers, however, in exhibiting a maximum percentage of fixations on the pattern for the medium level problems and a regression to a lesser percentage for the difficult problems, appear to execute a problem solving method characterized by a simple integration of past experience and the immediate solution of the problem. Their method seemingly cannot be intensified beyond the medium level problems. This regressive nature of the low performers' method in comparison with the progressive nature of the high performers' method may be a significant differential for discriminating problem solving attacks.

The percentages of fixations on the patterns and options were further analyzed by computing the mean percentages of consecutive fixations. The first set of consecutive fixations on the pattern may indicate on the subjects' first attempt to establish the rule. Whereas, the first set of consecutive fixations on the options may indicate the subjects first attempt to find a solution among the alternatives. The second and third sets of consective fixations may be second and third attempts to establish the rule or a verification of the first answer. These data were compiled in Table II.

On Number Series for high and low performers in the first examination the mean percentage of consecutive fixations on the patterns (P) was four to five times more than on the options (O). In the second examination both high and low performers fixated consecutively for the average percentage that was two to three times more on the patterns than on the options. While in the third examination, both high and low performers fixated consecutively for a more nearly equal mean percentage on patterns and options. High performers in the first and second examination fixated consecutively on the patterns for about as great a mean percentage of the total number of fixations as the low performers. But in the third examination of the patterns the low performers fixated consecutively for a mean percentage that was three times greater than that of the high performers. Similarly, in the first and second examinations of the options the high performers fixated consecutively for about as great a mean percentage as the low performers. Again, in the third examination of the options, the low performers fixated consecutively for a mean percentage of the total number of fixations that was seven times larger than the high performers' mean percentage. Therefore, in summation the high performers may be said to progress with a more thorough pattern analysis, as indicated by the greater mean percentage of consecutive fixations; i.e., the greater mean percentage of fixations on the pattern implied a more complete observation. The low performers appeared to be satisfied with a less complete analysis, and they looked to the options for clues to the answer; i.e., looking for a specific answer among the options would not require as many fixations as attempting to eliminate four options. Since in the third examination of the patterns the low performers fixated consecutively for a mean percentage of the total fixations that was three times larger than the high performers percentage, two different processes may have been oc-

TABLE I

EYE MOVEMENT MEASURES FOR LOW AND HIGH PERFORMERS ON NUMBER SERIES AND FIGURE ANALOGIES

	Number Series		Figure Analogies	
	Low	High	Low	High
Mean No. Fixations per Problem	37	21	21	18
Mean No. Regressions per Problem	29	8	9	7
Total Dur. Fixation 1/30 Sec.	444	287	244	195
Total Dur. Regression 1/30 Sec.	186	121	98	80
Mean No. Correct Answers	5.4	5.9	4.5	6.2

TABLE II

MEAN PERCENTAGE OF CONSECUTIVE FIXATIONS ON PATTERNS AND OPTIONS

Mean Percentage of Consecutive Fixations	Number Series				Figure Analogies			
	Low		High		Low		High	
	P%	0%	P%	0%	P%	0%	P%	0%
I Examination	45.0	10.3	60.6	12.0	24.4	21.1	39.9	25.5
II Examination	17.3	7.5	17.0	6.1	10.5	13.4	10.2	9.5
III Examination	7.9	7.0	2.5	1.1	7.0	7.4	5.4	3.1

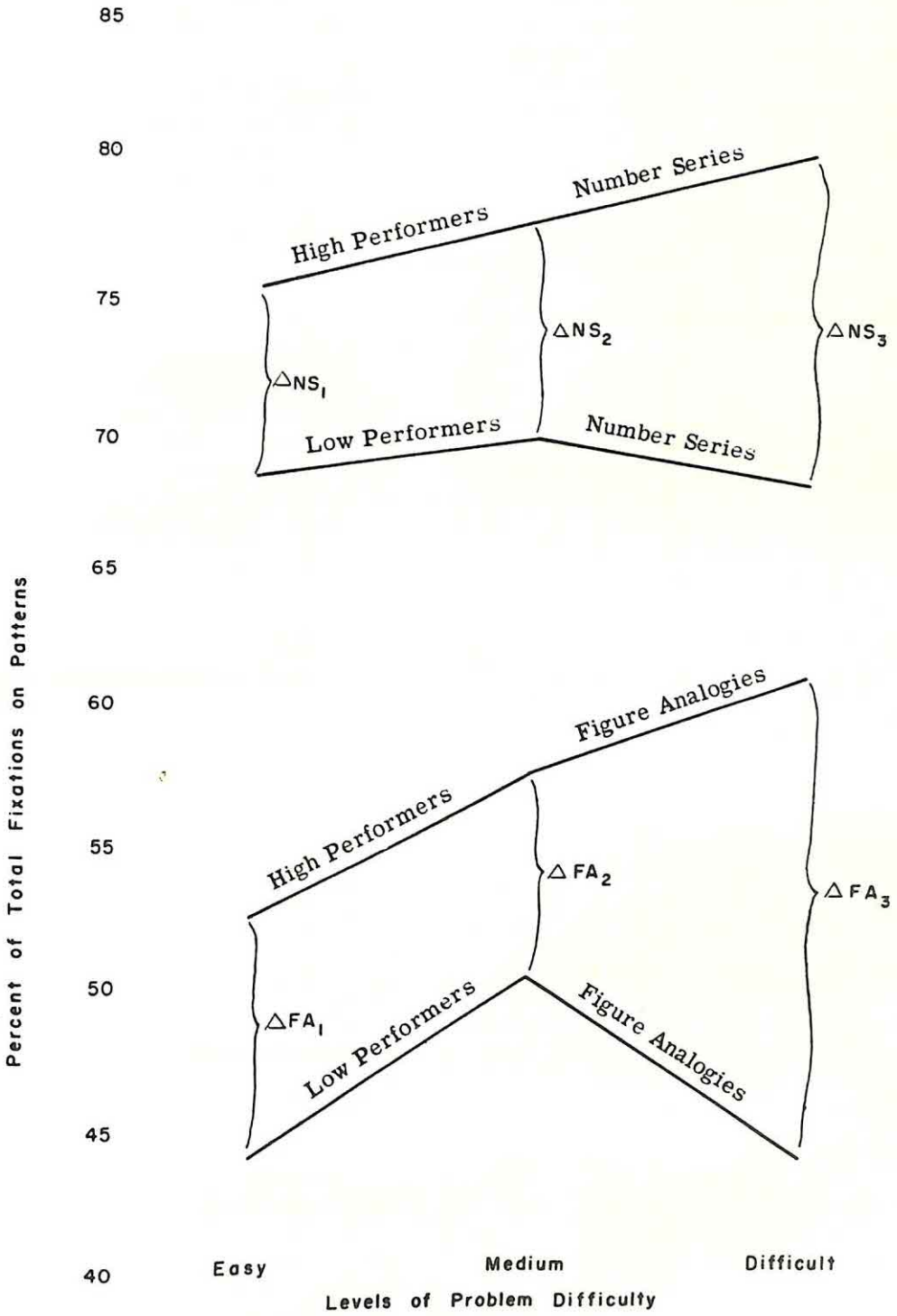


Figure 1. Percent of Total Fixations on the Pattern of the Problems as a Function of the Level of Difficulty

TABLE III

MEAN FREQUENCY OF ASCRIPTION OF RATING SCALE ITEMS DISCRIMINATING HIGH AND LOW PERFORMERS

Intervals for Mean Fre. of Ascription of Items	Rating Items of a Differential Mean Frequency for:			
	Low F. A.	High F. A.	Low N. S.	High N. S.
0 - 3.9	2 (3.0)		2 (2.8)	
4 - 6.9	9c (6.4)	2 (5.7), 3 (4.1)		2 (5.5)
7 - 10	3 (7.0)	9c (7.1)		

curring. The low performers appeared to be still searching for the rule; the high performers may have been verifying or correcting minor errors.

On Figure Analogies low performers fixated consecutively in all three examinations on the patterns for the same mean percentage as on the options. Contrastingly, the high performers scored a greater mean percentage of consecutive fixations on the patterns than on the options in the first examination, yet in the second and third examinations, they scored an almost equal mean percentage on the patterns and on the options, therein duplicating in part the results of the low performers. Similar to the processes suggested by the Number Series data, the high performers' first examination of the Figure Analogies may be interpreted as a more thorough pattern analysis than the low performers'. But there is a contrast of possible significance between the distribution of time in the third examination for the Figure Analogies and Number Series data; i.e., the difference between the mean percentage of consecutive fixations on the patterns of high and low performers was not as great as the same differential in the Number Series data. Moreover the same differential in the options in the Figure Analogies data is less than in the Number Series data.

In rating the verbal responses, the mean frequency of the ascription of items to the four groups was computed. Specifically, the mean frequencies, zero to three, four to six, and seven to ten inclusive, represented infrequent, moderately frequent, and very frequent respective ascriptions. From this analysis those items rated with a mean frequency different (in terms of the three categories above) for high and low performers on Figure Analogies were found to be the following: (item 2) identifying differences only, (item 3) identifying similarities and differences, and (item 9c) answering with the correct solution with no error.

Data for items two and three (Table II) strengthen the evaluation (from percentage of fixation data in Figure 1) of the high performers' method as analytic, wherein they tend to examine only differences. Notation of similarities would be less effective and more divertive. This indicates a greater purposefulness in the high performers' method of problem solving than in that of the low performers.

The finding that the item 9c, answering with the correct solution with no errors, discriminated only between the low and high performers on Figure Analogies is corroborated by the eye movement data (Table I). From these, the differential in mean numbers of correct answers was found to be greater for the low and high performers on Figure Analogies than on Number Series. This indicates a power differential for the low and high performers.

The finding that there were few rating items with a differential mean frequency for low and high performers, as yielded by this scale, may be attributable to the time factor in the administration of the test from which the selection of the subjects was made. For example, the effects of these variable factors may have largely contributed to concealing the low performers' solution as characteristically eliminative and that of the high performers as characteristically analytic.

Summary and Conclusions

This experiment was conducted to disclose the differential problem solving processes employed by high and low performers, respectively, in the solution of Number Series and Figure Analogies problems from the ACE Psychological Examination. A sample of 31 college students, from two classes in Introductory Educational Psychology, who had scored less than six or more than 19 Number Series items, or less than eight or more than 19 Figure Analogies items on either of the above mentioned sub-tests in which their performance was extreme. This examination consisted of individual, laboratory testing, first before the eye movement, and second with the use of an audograph for verbal recording. Eye movement measures were compiled from the developed film, and the verbal recordings were evaluated on a dichotomous rating scale.

Relevant to the questions investigated, the results suggest the following conclusions:

1. High performers exhibit fewer fixations and regressions than low performers in solving Number Series problems. High and low performers employed a similar number of fixations and regressions in Figure Analogies problems.

2. High performers have a total duration of fixations and regressions which is less than the low performers' in solving Number Series problems. For Figure Analogies problems the duration of fixations and regressions is not very different for the two groups.

3. The high performers fixated more on the pattern than on the options for all problems of the three levels of difficulty. Moreover, their percentage of fixations on the pattern increased with increasing difficulty level of the patterns. For low performers, the former finding was true only on the Number Series problems. The percentage of their fixations on the pattern increased only from the easy to the medium level problems then decreased for the difficult problems. This was indicated in their performance on both Figure Analogies and Number Series problems.

4. There was no substantiation in either the

eye movement or verbal-recorded data that high performers shift more readily from one arithmetic process required to solve a problem to another process for a subsequent problem.

The verification of the questions under investigation may have been limited by the assumption that eye movement responses are in part symptomatic of thought processes. This may not be true. Rather thought processes may be a "delayed-reaction-expression" of the eye movement response (Thought I is concurrent in time with Fixation 2); i.e., the eye movement versus the thought response may be analogous to the eye movement versus the voice span record.

REFERENCES

1. American Council on Education Psychological Examination (Washington, D. C.: American Council on Education, 744 Jackson Place).
2. Anselmo, V. J. An Eye Movement Study of Number Series Completions, Unpublished Master's Thesis, University of California, 1940.
3. Gilbert, L. C. "An Experimental Investigation of Eye Movements in Learning to Spell Words," Psychological Monographs, XLIII (1932), pp. 1-81.
4. Greening, C. P. Differential Factors in the Solution of Figure Analogies Problems by High and Low Achieving Individuals, Unpublished Master's Thesis, University of California, 1948.

ACADEMIC ATTRITION OF ENGINEERING TRANSFER STUDENTS

J. STANLEY AHMANN
Cornell University

PERIODICALLY some of the careful observers of the higher education scene express increasing concern about the relatively high rates of academic attrition found at most institutions. In spite of any benefits which a student may receive from a contact with a college or university, be it as little as a single semester, the opinion is stated that the failure of sizable percentages of students to graduate has resulted in an unnecessary dissipation of energies and finances. Were a rechanneling of these efforts possible, appreciable gains to both the institutions and the students concerned are envisioned.

An answer to the problem would be, of course, a more careful screening of applicants requesting entrance to engineering curricula. Efforts to find those characteristics which are highly related to academic success in engineering have been numerous (8). The predictive usefulness of the high school grade-point average (9), scholastic aptitude tests (2, 5, 11), other aptitude tests (10), reading tests (3), interest tests (7), and personality scales (6), individually and in combination, has been investigated. In many instances the results have been promising, even though incapable of offering a near perfect selection scheme.

In the case of engineering colleges in which sizable numbers of students enter as transfer students, the problem of selecting the most promising students is further complicated. At the Iowa State College, for example, estimates have been made that as many as 40 percent of the entering engineering students at the beginning of a fall term had received college credit from other institutions of higher education. A study (1) of the transfer students entering the Engineering Division of this college during the 1946-47, 1947-48, and 1948-49 academic years revealed that most students (80%) had attended only one college prior to enrolling at the Iowa State College, and that the college was usually located in Iowa or an adjacent state and enrolled less than 2500 students. Furthermore, of the 804 students included in the study, only 246, or 31%, graduated in engineering. The remaining 558 either failed, transferred to non-engineering curricula at the Iowa State College, transferred to other institutions of higher education, or dropped from college for miscellaneous personal reasons. Even though a few of the 558 may have been academically successful elsewhere, they can be properly classified as attrition students in the eyes of the engineering faculty.

Although the foregoing study investigated the relationships between a series of numerical variables and the tendency to graduate in engineering, no attempt was made to study the possible influence of non-numerical characteristics on this criterion. An extension of the study, therefore, seemed in order.

On the basis of a preliminary examination of the data available, one of the non-numerical factors which seemed to warrant examination was the type of institution first attended by the transfer student. Although this factor was but one of many potentially influencial factors, indications were found that it was possibly more influential than most. Therefore, the following report is restricted for the most part to the single consideration of whether the type of college at which a transfer student first matriculated affected his tendency to graduate in engineering at the Iowa State College.

For purposes of classification, the engineering transfer students were considered to have matriculated for the first time at one of two different types of institution, either one offering only a two-year program or one offering more than a two-year program. The hypothesis was then posed as to whether, with respect to transfer students entering engineering curricula at the Iowa State College, those who first matriculated at institutions offering only a two-year program differed from those who first matriculated at institutions offering more than a two-year program in terms of tendency to graduate in engineering.

A random sample of 256 male engineering transfer students was selected from the 804 students included in the original study. This sample was so drawn that students having matriculated at both types of institutions were equally represented. Furthermore, since earlier research (4) demonstrated that students who were veterans of World War II tended to surpass non-veteran students in academic achievement, the sample was further sub-divided on that basis, thus yielding four subgroups with 64 cases included in each subgroup. In Table I is shown the number of students in each subgroup who graduated in engineering.

Inspection of this table revealed that, when individual differences in academic aptitude were ignored, sizable differences in tendency to graduate existed. The students first matriculating

TABLE I

TENDENCY TO GRADUATE IN ENGINEERING OF 256 TRANSFER STUDENTS

Type of College		Veteran Status: Yes Grad.	Yes Attrition	No Grad.	No Attrition	Total Grad.	Total Attrition
Two-Year Program Only	k	19	45	12	52	31	97
	%	29.7	70.3	18.8	81.2	24.2	75.8
More Than Two-Year Program	k	28	36	17	47	45	83
	%	43.8	56.2	26.6	73.4	35.2	64.8
Total	k	47	81	29	99	76	180
	%	36.7	63.3	22.7	77.3	29.7	70.3

at an institution with more than a two-year program seemed to graduate in distinctly greater numbers than those who first matriculated at institutions offering only a two-year program. Also, veteran students obviously surpassed non-veteran students with respect to this criterion. Of the four subgroups, the veteran students first matriculating at institutions offering more than a two-year program seemed definitely to excell.

To test the significance of the differences in tendency to graduate in engineering, an analysis of variance can be computed provided an assumption is made concerning the nature of the graduation-attrition dichotomy. In this case, the assumption was made that the tendency to graduate in engineering was a single normally distributed variable and was no more sensitively measured than by the graduation-attrition classification. This assumption does, therefore, underlie all of the procedures and interpretations made in the following paragraphs.

The steps followed in the computation of the analysis of variance were the same as those reported by Wert, Neidt, and Ahmann (12). According to this procedure sums of squares were found for the main effects of type of college and veterans status as well as for the interaction by designating $\frac{z}{p}$ as the value to be assigned to each member of the graduation groups, and $\frac{z}{q}$ as the value to be assigned to each member of the attrition groups. The quantities p and q are the proportions of the total sample of 256 students who graduated and did not graduate in engineering respectively. The value z is the height of the ordinate dividing the normal curve of unit area in p and q parts. The entries in the analysis of variance table were then found in somewhat the same manner as in the problems in which a numerical criterion is present. The results are summarized in Table II.

The F-value for the type of college main effect failed to meet significance at the 5% level by a very slight amount. The conclusion, therefore, was considered to be in doubt. The possibility remained that those transfer students who first matriculated at institutions offering only a two-year program did, as a group, experience greater difficulty in graduating because of that fact. In the case of the remaining two F-values, the significance of that for the veteran status main effect and the non-significance of the value for the interaction were not surprising.

In the foregoing analysis any individual differences in studentship which might have influenced tendency to graduate in engineering on the part of transfer students have been ignored. To investigate the possible influence of type of college on transfer students' tendency to graduate in engineering, an analysis corresponding closely to the analysis of covariance was needed in which individual differences in studentship were controlled.

The quantitative raw scores on the American Council on Education Psychological Examination and the high school grade-point averages were available for all students and were used as indicators of studentship. The latter values were tabulated on an A, B, C, D, and F basis, and then converted to a 4, 3, 2, 1, and 0 basis. The mean values of both variables are shown in Table III.

In all four subgroups the difference between the graduation group and the corresponding attrition group was striking with respect to caliber of studentship as represented by these two variables. Differences between the means of the quantitative scores were often as great as 10 points and once almost 20 points. Differences between the means of the high school grade-point averages were usually 0.2 or 0.3. In every instance the graduation group surpassed the attrition group.

Of additional importance, even though not included as such in Table III, was the fact that, as a group, the transfer student representing the one type of college differed from those representing the other in the following manner. The mean quantitative score and mean high school grade-point average for the transfer students first matriculated at an institution with only a two-year program were 61.6 and 2.62 respectively. The corresponding values for the transfer students who first matriculated at institutions offering more than a two-year program were 64.4 and 2.68. In terms of these two variables, therefore, the institutions offering the longer program tended to attract the better students.

In order to control on the individual differences in studentship as represented by these two measures, the analysis of variance shown in Table II was expanded into a variation of the ordinary analysis of covariance. This variation, although much the same as the original analysis of covariance, employed modified discriminant functions (12) in place of the regression equations. The discriminant functions were of the same number and type as the regression equations used in covariance analysis and served much the same function.

The results of the analysis are shown in Table IV. It should be noted that the proportions of the individual differences in graduation tendency that could be explained by variations in the quantitative scores and the high school grade-point averages were computed, and were then expressed as the proportion of the variance representing individual differences in graduation tendency not associated with variations in the two numerical variables. The resulting values were 0.8458, 0.8588, 0.8508, and 0.8510. With these known proportions, it was possible to re-

TABLE II

ANALYSIS OF VARIANCE OF TENDENCY TO GRADUATE IN ENGINEERING

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F
Type of College	1	2.1026	2.1026	3.65
Veteran Status	1	3.4747	3.4747	6.04
Interaction	1	0.1727	0.1727	0.30
Within	255	146.7524	0.5755	

TABLE III

MEANS OF QUANTITATIVE FAW SCORES (X_1) AND HIGH SCHOOL GRADE-POINT AVERAGE (X_2)

Type of College		Veteran Status: Yes, Grad.	Yes, Attrition	No, Grad.	No, Attrition	Total, Grad.	Total, Attrition
Two-Year Program Only	$\bar{X}_1$	72.9	53.7	75.1	61.2	73.7	57.7
	$\bar{X}_2$	2.79	2.46	3.20	2.55	2.95	2.51
	k	19	45	12	52	31	97
More Than Two-Year Program	$\bar{X}_1$	71.6	60.5	66.6	62.4	69.7	61.6
	$\bar{X}_2$	2.86	2.52	2.88	2.64	2.86	2.58
	k	28	36	17	47	45	83
Total	$\bar{X}_1$	72.1	56.7	70.1	61.8	71.4	59.5
	$\bar{X}_2$	2.83	2.49	3.01	2.59	2.90	2.55
	N	47	81	29	99	76	180

TABLE IV

ANALYSIS OF COVARIANCE OF TENDENCY TO GRADUATE IN ENGINEERING

Source of Variation	"Within" Plus Source of Variation					Source of Variation Alone		
	Unadjusted		Proportion Not Associated	Adjusted				
	d.f.	S.S.		d.f.	S.S.	d.f.	M.S.	F
Type of College	256	148.8550	0.8458	254	125.9016	1	1.0153	2.06
Veteran Status	256	150.2271	0.8588	254	129.0150	1	4.1287	8.36
Interaction	256	146.9251	0.8508	254	125.0039	1	0.1176	0.24
Within Alone	255	146.7524	0.8510	253	124.8863		0.4936	

turn to the information assembled in the analysis of variance shown in Table II, and remove, as in the common analysis of covariance, any allowance which need be made because of individual differences between the groups on the control factors. The adjusted sums of squares were converted to mean squares and the F-values computed in the usual manner.

The F-value for the type of college main effect failed to reach the 5% level of significance. Therefore, insofar as the quantitative scores and high school grade-point averages controlled individual differences in studentship, and no other factors contributed a bias, no significant differences have been found in tendency to graduate in engineering at the Iowa State College between transfer students first matriculating at an institution offering only a two-year program and those transfer students first matriculating at institutions offering more than a two-year program. It was concluded that the possibility of matriculating at an institution offering a broader program enhanced a transfer student's tendency to graduate in engineering, as suggested in the analysis in Table II, disappeared when individual differences in studentship were considered. As suggested in an earlier paragraph, it can be inferred that it was not the type of program offered as such which caused transfer students from two-year programs to tend to have greater difficulty in graduating in engineering, but rather that such institutions seemed to enroll less talented students in this instance.

REFERENCES

1. Ahmann, J. Stanley. Prediction of Achievement of Iowa State College Engineering Students Having Transferred from Other Institutions, Unpublished Ph. D. Dissertation, Iowa State College Library, Ames, Iowa, 1951.
2. Berdie, R. F. and Sutter, N. A. "Predicting Success of Engineering Students," Journal of Educational Psychology, XLI (March 1950), pp. 184-190.
3. Feder, D. D. and Adler, D. L. "Predicting the Scholastic Achievement of Engineering Students," Journal of Engineering Education, XXIX (January 1939), pp. 380-385.
4. Gowan, A. M. Unique Characteristics of Freshman Veterans at the Iowa State College with Administrative Implications, Unpublished Ph. D. Dissertation, Iowa State College Library, Ames, Iowa, 1947.
5. McClanahan, W. R. and Morgan, D. H. "Use of Standard Tests in Counseling Engineering Students in College," Journal of Educational Psychology, XXXIX (December 1948), pp. 491-501.
6. MacRae, J. M. Usefulness of the Minnesota Personality Scale for Predicting Achievement of Freshman Engineering Students, Unpublished M. S. Thesis, Iowa State College, Ames, Iowa, 1949.
7. Minor, W. T. Usefulness of the Kuder Preference Record for Predicting Academic Success of Iowa State College Engineering Freshmen, Unpublished M. S. Thesis, Iowa State College Library, Ames, Iowa, 1947.
8. Moore, J. E. "A Decade of Attempts to Predict Scholastic Success in Engineering Schools," Occupations, XXVIII (November 1949), pp. 92-96.
9. Pierson, G. A. Jr. "School Marks and Success in Engineering," Educational and Psychological Measurements, VII (Autumn 1947), pp. 612-617.
10. Treumann, M. J. and Sullivan, B. A. "Use of the Engineering and Physical Science Aptitude Test as a Predictor of Academic Achievement of Freshmen Engineering Students," Journal of Educational Research, XLIII (October 1949), pp. 120-133.
11. Vaughn, K. W. "The Yale Scholastic Aptitude Tests as Predictors of Success in College of Engineering," Journal of Engineering Education, XXXIV (April 1944), pp. 572-582.
12. Wert, James, E. and others. Statistical Methods in Education and Psychology (New York: Appleton-Century-Crofts, Inc., 1954), Chs. 15 and 19.

COLLEGE LEVEL STUDY SKILLS PROGRAMS: SOME OBSERVATIONS

WALTER S. BLAKE, Jr.
University of Maryland

COLLEGE-LEVEL study skills programs are becoming more numerous. Twenty-four institutions are planning such programs for the near future. Institutions of higher learning are enrolling anywhere from seven to 1400 students in their programs in the United States and Possessions, and all programs in which evaluations have been undertaken report favorable results. However, most of the programs seem to resemble "Topsy" somewhat—they just "growed up" without the benefit of the experiences of others by virtue of the fact that the experiences of others in this field have not been reported in the literature in any appreciable measure.

The University of Maryland began a program in 1947, and it, too, grew out of experimentation at Maryland largely rather than as a result of the experiences of workers in other programs. However, a study was undertaken in 1953 to survey and evaluate both the program at the University of Maryland and other programs in operation throughout the United States and Possessions in order to improve the program at the University of Maryland in the light of the findings.* The workers in the University of Maryland program feel that at least part of what they found out could benefit workers in other programs, and so the following highlights of the findings and recommendations from the study are presented in the hope that the many program workers and their students will find the information useful to them.

1. Most programs offer services to a limited segment of the school population. Forty-two and two-tenths percent admit voluntary and referral students (probationers, etc.); 40% admit only voluntary students, and 11.1% require all freshmen to enroll (with a few taking voluntary students as well). Six percent did not report in this area. The wide variation of admission policies is surprising since the consensus is that any study skills program is composed of guidance services which should be available to the entire student body if the program is to attain its greatest effectiveness. All entering freshmen should be assigned to a program designed to indoctrinate them to the life on campus plus the minimal skills needed to achieve their goals at college and afterward; and the services of the program (tutorial, remedial reading, study skills and reading courses, counseling etc.,) should be open to all students on campus who feel a need for such services.

2. The "remedial" aura still surrounds and plagues study skills programs, in general. The remedial phase(s) of most programs take precedence over the preventative phases, with the result that very few schools make provisions for helping any students other than those who must be helped. The "average" student is obliged to struggle along without assistance until he, or some faculty member, notices that he is about to fail out of college, at which time "remedial" measures may be taken (if it is not already too late). In most institutions where no required program for freshmen is offered, faculty referrals and self-referrals are the only means available to help prevent academic failure and social maladjustment.

3. Program-planning with students is conpiculously lacking in many of the programs surveyed. Small staffs and insufficient operating funds usually account for this; yet the absence of student-faculty planning is a serious shortcoming, nonetheless, in programs of this kind. The types and extent of services offered should be the result of student-faculty planning, based upon research findings. One way to help insure student participation in the program is to incorporate student-faculty planning as a part of the program itself. Written student evaluations, soliciting student suggestions, interviews with students, consultation with student government leaders, and regularly scheduled student-faculty meetings are useful methods. The main point here is this: faculty-seen needs are not necessarily student-seen needs—a well-known fact often overlooked. It is recognized that a well-trained faculty might know more about what students need than the students themselves, yet this obviously does not guarantee student acceptance of a program planned entirely by faculty members. Student-faculty planning might well be termed a "calculated risk" in the study skills area; but it seems no less essential than in any other situation where democratic procedures seem likely to produce the best results.

*This article is based upon a doctoral study completed by the author entitled: A Survey and Evaluation of Study Skills Programs at the College Level in the United States and Possessions, University of Maryland, 1953.

4. Research is being done neither in the minimal quantity necessary nor in the areas where it is most needed. The quantity of research needed will necessarily be governed by needs of individual programs; but every program needs research of the kind which will indicate (1) whether the program is achieving set goals, and (2) what needs to be done to improve the program. While it is true that program workers spend most of their time giving service (as do most people in the various branches of the teaching profession), it is equally true that a part of every worker's time needs to be devoted to research in the program if the program is to be successful, and if the workers are to have confidence in the program itself as well as their part in the program. Research is needed particularly in these areas: program evaluation, program improvement, and validation of diagnostic instruments.

5. Over half (51.1%) of the programs surveyed do not give academic credit for participation in the formalized parts (classes in study skills and reading, mainly) of the programs. Credit is "expected" by college students for work done under the auspices of the institution out of habit and tradition. Good or bad, it is nonetheless true that college credit is a motivating factor with college students—perhaps the most important single motivating factor. It is also true that student initiative is important to any student's success or failure in meeting or solving his problems. Therefore, it seems important to make the process of problem-solving in any group guidance situation as profitable as possible to students in order to nurture initiative. Some workers who do not give academic credit feel that some of the services rendered and some of the course materials and techniques used are not "college level" in terms of the conventional college-level courses. While such may actually be the case in many programs, the failure to grant some credit for work accomplished may doom good programs to ineffectuality, no matter how fine such programs may be potentially.

6. Study skills programs need workers trained to work in study skills programs. At present, nearly all workers are educators, psychologists, or other kinds of specialists not necessarily trained to be workers in study skills programs. Workers having majored in areas such as education and psychology might have some of the general qualifications needed (like the desire to work with students); but workers could have the special qualifications needed only by chance. For example, educators do not usually learn psychology in their curriculum, and psychologists do not learn teaching methods; yet both psychology and teaching methods are acknowledged to be two of the important special qualifications desirable for program workers by program workers themselves. Only one institution, out of the many contacted in the survey, offers a training program specifically for study skills program workers, yet hundreds of persons are now employed in such programs, and 24 institutions plan such programs for the future.

7. Study skills programs are not publicized adequately, as a rule — indeed, some are kept on a "confidential" basis among staff members. The reticence on the part of the program workers to make their services known does a disservice to the student body and also prevents the programs from reaching their maximum level of effectiveness. Evidence points to frugal financing of such programs as the basic reason for curtailed services as well as lack of publicity about services offered; but it seems certain that a program designed to help students cannot be kept secret from students and at the same time serve their needs. The publicizing of programs need not be of the conventional advertising variety, of course; but the program should be made known to all students through written notices concerning services, hours, etc., articles in the campus literature which will reach and be read by both students and faculty, and any other device available to workers. The students and faculty who have received satisfactory service provided by the program will, of course, be the best publicity mediums, once the program has been operating long enough to become known on the campus.